IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2019, accepted April 21, 2019, date of publication April 26, 2019, date of current version May 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913442

Towards Accurate High Resolution Satellite Image Semantic Segmentation

MING WU, CHUANG ZHANG, (Member, IEEE), JIAMING LIU, LICHEN ZHOU, AND XIAOQI LI
Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China
Corresponding authors: Ming Wu (wuming@bupt.edu.cn) and Chuang Zhang (zhangchuang@bupt.edu.cn)

This work was supported by the National Natural Science Fund under Grant 61773071.

ABSTRACT Satellite image semantic segmentation, including extracting road, detecting building, and identifying land cover types, is essential for sustainable development, agriculture, forestry, urban planning, and climate change research. Nevertheless, it is still unclear how to develop a refined semantic segmentation model in an efficient and elegant way. In this paper, we propose attention dilation-LinkNet (AD-LinkNet) neural network that adopts encoder–decoder structure, serial–parallel combination dilated convolution, channel-wise attention mechanism, and pretrained encoder for semantic segmentation. Serial–parallel combination dilated convolution enlarges receptive field as well as assemble multi-scale features for multi-scale objects, such as long-span road and small pool. The channel-wise attention mechanism is designed to advantage the context information in the satellite image. The experimental results on road extraction and surface classification data sets prove that the AD-LinkNet shows a significant effect on improving the segmentation accuracy. We defeated the D-Linknet algorithm that won the first place in the CVPR 2018 DeepGlobe road extraction competition.

INDEX TERMS Satellite image, semantic segmentation, AD-LinkNet, dilated convolution, channel-wise attention.

I. INTRODUCTION

Satellite image semantic segmentation is a pixel-wise classification task for a satellite image. Satellite images are gaining attention from the community for map composition, population analysis, effective precision agriculture, and autonomous driving tasks because satellite imagery contains more structured and uniform data compared to traditional images [1]. Understanding satellite image including extracting road, detecting building, and identifying land cover types are essential for sustainable development, agriculture, forestry, urban planning and climate change research. Road extraction, building detection and land cover classification are based on semantic segmentation task.

Image semantic segmentation has gained remarkable improvement with the development of fully convolutional neural networks. compared with the general semantic segmentation tasks, the challenges of high-resolution sub-meter satellite image semantic segmentation are to produce finer predictions for every pixel in the large-scale image. There are strong differences between satellite imagery and everyday pictures, such as PASCAL VOC2012 [2] and Microsoft COCO [3]. Satellite imagery assumes a bird's view acquisition, thus objects lie within a flat 2D plane and every pixel in satellite images has a semantic meaning. However, the PASCAL VOC2012 dataset are assume a human-level point of view and mainly comprised of meaningless background with a few foreground objects of interest [4].

LinkNet [5] is an efficient semantic segmentation neural network which takes the advantages of skip connections, residual block [6] and encoder-decoder architecture. The original LinkNet uses ResNet18 as its encoder, which is a pretty light but outperforming network. LinkNet has shown high precision on several benchmarks [7] and it runs pretty fast. D-LinkNet uses LinkNet [8] with pretrained encoder as its backbone and has additional dilated convolution layers in the central part.

Satellite image contains multi-scale objects: main road stretching across a whole image (see Figure 1 (a), small farmland inlaying an urban (see Figure 1 (b). Dilated convolution is a useful kernel to adjust receptive fields of feature points without decreasing the resolution of feature maps. It has two types, cascade mode like [9] and parallel mode like [10]. We add shortcuts to the series dilated convolution, which makes the series structure expand into a series-parallel structure.

The associate editor coordinating the review of this manuscript and approving it for publication was Yi-Zhe Song.

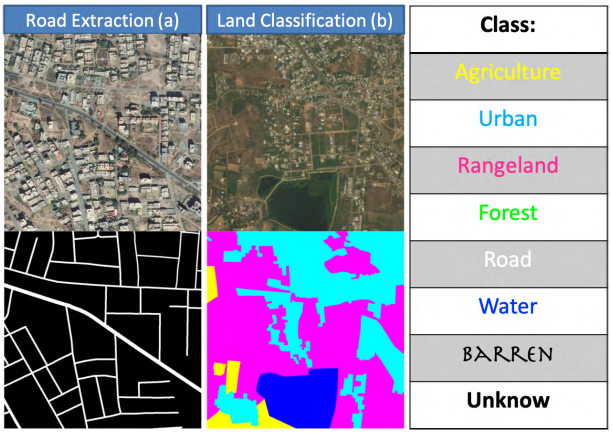


FIGURE 1. Example diagram of task introduction.

Satellite image contains rich context information. For example, ''roads'' generally cannot directly pass through ''buildings'', We proposed AD-LinkNet to leverage context information to benefit satellite image semantic segmentation task by introducing channel-wise attention [11].

The size of annotated satellite image datasets are small. Transfer learning is a useful method that can directly improve network performance in most situation [12], especially when the training data is limited. In semantic segmentation field, initializing encoders with ImageNet [13] pretrained weights has shown promising results [10], [14]. We initialize AD-LinkNet encoder with ImageNet pretrained weights.

Data augmentation is essential to prevent overfitting. We augment datasets in an ambitious way, including horizontal flip, vertical flip, diagonal flip, ambitious colour jittering, image shifting, scaling.

We used the road extraction and land cover classification datasets of CVPR2018 DeepGlobe Challenge to examine the effect of AD-LinkNet, and won the 1st places in the road extraction task, and got the top ten places in the land classification task. The main contributions of our work are as follows:

- We analyze the effectiveness of several properties for satellite image semantic segmentation, and reveal how to leverage them to benefit the satellite image semantic segmentation task.
- We design a simple yet effective AD-LinkNet structure by leveraging the useful properties to conduct satellite image semantic segmentation in a simple and efficient way.
- Our AD-LinkNet brings a significant performance boost to satellite image semantic segmentation: road extraction task, outperforming the current state-of-the-art method.
- Our code is available, which can serve as a solid baseline for the future research in satellite image semantic segmentation such as road extraction and land cover classification.

II. BACKGROUND

In this section we will introduce the state of the art approaches for road extraction and land classification tasks. Meanwhile, we will introduce the background knowledge of important modules in AD-LinkNet, including the dilated convolution and Attention mechanism. Finally, this section will introduce the knowledge of our universal approaches to achieving great results in the competition, which include transfer learning and data augmentation.

A. SEMANTIC SEGMENTATION OF SATELLITE IMAGE

Satellite image segmentation, used to locate objects and boundaries in images (straight lines, curves, etc.), refers to the division of a digital image into multiple pixel sets. More precisely, image segmentation is the process of assigning a label to each pixel in an image, same-labeled pixels with same characteristic [15].

There is a long tradition of using computer vision techniques for satellite image understanding [16], [17]. Historically, satellite imagery was typically lower-resolution, from a strictly top- down view, and with a diversity of spectral bands. The segmentation method based on deep learning emerged in recent years. Since the fully convolutional network (FCN) [18] has shown numerous improvements in semantic segmentation, many researchers [19]–[21] have made efforts based on the FCN. The network model designed in this paper are based on the FCN. And then, Unet [22] uses Transposed-conv [23] as its upsampling structure on the basis of FCN, connects the features of the network Encoder part to the Decoder part, and combines low-lever information with high-level information (Such an hourglass & shortcut connection structure is called U-shape). Volpi and Tuia [24] also proposed to use an subsample-upsample architecture in satellite semantic segmentation task, which like U-shape structure. This paper selects Unet as one of the baselines. At the same time, LinkNet [5] with ResNet [6] as Backbone is also one of the baselines for this project. LinkNet also uses the U-shape structure and replaces the convolution structure of each level of its Encoder and Decoder with a res-block. This network has a rich shortcut, which is more conducive to transmitting shallow information to deeper layers of the network. And we have used LinkNet34 as the basic module of our previous network model(D-LinkNet34) [8].

B. DILATED CONVOLUTION

The dilated convolution can be used as a general convolution for weighting operation, and at the same time has the function of pooling layers to multiply the receptive field. As the general convolution layers, dilated convolutions can be stacked layer by layer to form a series structure. The dilated conv is to add the ''dilated rate'' to the last convolution layers of the pre-trained classification network when performing the transfer learning. Double the dilated rate of the convolution for each pooling layer removed and maintain the same dilated rate under the same feature resolution. OverFeat [25] first

applied dilated convolution in deep learning to deal with object localization problem. Deeplab [10] first named this convolution structure and converted the series structure into a parallel structure. Dilated ConvNet [9] described in detail the series convolution and how it achieves the exponential expansion of the receptive field. Dilated ConvNet [9] and DRN [26] used the dilated convolution to amplify the network receptive field. This kind of dilated convolutions amplify the original classification feature map by multiple times, which constitutes a Large-scale feature maps with rich spatial information and suitable for semantic segmentation.

C. ATTENTION MECHANISM

The Attention mechanism has a great improvement on the sequence learning task. In the encoder-decoder framework, the source data sequence is weighted by adding the Attention mechanism in the encoder. Or the Attention mechanism is introduced at the decoder, and the weighted change of the target data can effectively improve the system performance of the sequence pair sequence in the natural mode [27]. There is a problem with the LSTM/RNN model of the traditional encoder-decoder structure: it is encoded into a fixed-length vector representation regardless of the input length, which makes the model poorly learn for long input sequences. The Attention mechanism overcomes the above problem. The principle focus on the relevant corresponding information selectively in the input when the model is outputting.

Soft attention developed in recent work [19], [28] can be trained end-to-end for convolutional network. Attention to scale [19] uses soft attention as a scale selection mechanism and gets state-of-the-art results in image segmentation task [18], [29]. In the semantic segmentation task, there is a problem with the convolutional neural network model: for large-scale image as input, the model is difficult to learn all the information. Especially for satellite images, segmentation targets often exist as small targets, and it is difficult for neural network to make accurate segmentation training. However, the Attention mechanism can process large-scale images before the model is predicted [27], making models better suited for large-scale image segmentation tasks.

D. TRANSFER LEARNING AND DATA AUGMENTATION

In the field of image semantic segmentation, the pre-trained data can come from image semantic segmentation data or image classification data. Be similar to the image classification task, image semantic segmentation also has the large and general object segmentation data set. These data sets can be used as a pre-trained semantic segmentation network, and then we can fine-tune the network on other data sets. However, the data set of image semantic segmentation often labels fewer types of objects (for example, COCO initially only labels 80 types of semantic labels). In this case, some methods choose to use a more extensive image classification dataset (ImageNet, etc.) to pre-train the network and then tune it on a large-scale semantic segmentation dataset, at last, final tuning on the target's segmented data set [30].

Data augmentation is very important for training deep networks. Even with large data sets, the use of reasonable data augmentation methods can still improve the performance of the network [6]. In the field of image semantic segmentation, the scarcity of data is more apparent. On some insufficient data sets, it is often necessary to use more ''radical'' data augmentation methods. The data augmentation methods are all performed during network training. In fact, in some scenarios where not require high real-time effect, some data augmentation methods used in training can be applied to the test data, which also lead to better test results.

E. DEEP UNET AND LINKNET

Unet and LinkNet are the basic modules of the baseline and AD-LinkNet for our experiment. So we introduce these two models in this section. This paper does not directly adopt the original Unet network model, but makes appropriate improvements to the original Unet, and makes it more suitable for the experimental requirements of the project. Unet is a segmentation network for medical tissue cell images. The central part has a small receptive field of 140*140 per feature point, which is not suitable for other tasks, and the input image size must be fixed at 572*572. However, the data set for the road extraction task is 1024*1024. The improved Unet differs from the original Unet in terms of the basic structure. The improved Deep Unet increases the Padding layer and the BatchNorm (BN) layer. The Padding layer allows the network to be maintained during convolution, and the BatchNorm layer allows the network to capture the distribution of the data set easily, which promotes the convergence of network. The basic structure of the original Conv-ReLU is extended to the structure of Padding-Conv-BatchNorm-ReLu. In this paper, we expand the four subsampling processes of the original Unet to seven, which increases the network depth and greatly increases the receptive field of the central network (to 1148*1148), making the network suitable for a variety of tasks.

LinkNet [5] is a variant of U-shape, which differs from Unet in two main points. Firstly, it replaces Unet's ordinary convolution structure with residual module (res-block). Secondly, it transforms Unet's deep and shallow feature synthesis method from ''stacking'' to ''adding''. Original LinkNet, which is one of the lightest ResNet, uses ResNet18 as its Encoder. Such LinkNet18 can guarantee both high accuracy and forward propagation efficiency of the network. In practice, different types of LinkNet can be obtained by transforming the Encoder part of LinkNet into ResNet with different depths and different representations. Therefore, the operational accuracy and efficiency can be weighed by adjusting the number of layers of Encoder. In the meantime, due to the fact that the Encoder of LinkNet maintains the same structure as ResNet, the pre-trained ResNet can be directly used as the Encoder of LinkNet. This kind of transfer learning makes LinkNet converge faster and has stronger

generalization ability. With five subsampling processes (four pooling and one step convolution), LinkNet's central characteristic resolution is higher than that of deep Unet.

III. AD-LINKNET (ATTENTION DILATION - LINKNET)

From the perspective of network structure evolution, according to the characteristics of image semantic segmentation, we propose a new refined segmentation network step by step, and finally propose AD-LinkNet which integrates the advantages of multiple networks and is based on our previous D-LinkNet34 [8]. Based on the inheritance of D-LinkNet's outstanding features, the AD-LinkNet adds a Series-parallel combination dilated convolution and an Attention mechanism in the network to form a refined semantic segmentation network. This article discusses AD-LinkNet's mechanism and framework, then compares its performance with D-LinkNet34 in satellite image processing task.

A. SERIES-PARALLEL COMBINATION DILATED CONVOLUTION

About the choice of dilated convolution, the original author of ResNet believes that, the validity of the residual structure(res-block) is derived from the ''identity mapping'' of the residual structure(res-block), which benefits the back-propagation of the network gradient as well as solves the gradient dissipation problem effectively [31]. However Sergey et al. proposed the wide residual network [32], stating that the residual network does not necessarily need to be so deep, while some networks with fewer layers can even surpass the performance of the deep residual network when using the residual structure(res-block). And Veit *et al.* [33] claimed that ''identity mapping'' may not be the reason for ResNet to improve network performance, it is due to the shortcut connection. Recently, Wu *et al.* [34] designed a model of ''residual module selection'', which can choose different residual modules (res-block) to pass data according to different input data.

In this paper, we use the characteristics of ''parallel expansion'' of the residual network, and use the short-cut connection to make the dilated convolution also form a structure which is series-parallel combination. This structure has the function of connecting the dilated convolution expansion network receptive field series, and also connecting the dilated convolution comprehensive multi-scale semantics parallel. This structure is the most crucial part in AD-LinkNet for network performance enhancement. Next, we will describe a series-parallel combination dilated convolution and reveal the advantages of this structure for feature fusion and receptive field augmentation.

A parallel dilated convolution structure allows the feature map to use a variety of convolutional structures with different dilated ratios, and then fuses the information of different branches by ''stacking'' to achieve multi-scale feature fusion. However, the parallel structure has the same depth for each branch, and each of them has only a single convolution layer. There is a certain similarity between each branches, so the diversity of features is lacked.

Inspired by the ''extension of the res-block to parallel'', we add a short-cut connection in the series dilated convolutions to form a dilated convolution of the res-block, which can be decomposed into the form of multiple branches. We places this structure in the central part of LinkNet, and proposes AD-LinkNet for refined segmentation tasks.

B. CHANNEL-WISE ATTENTION

We used SE-Net and SE-Loss in the model. For SE-Net [35], this global feature is used to make channel-wise attention to other branches of the network. This Attention mechanism enhances the usage of the effective feature layer by weighting the ''importance'' of different feature layers. For SE-Loss [36], the classification information is also incorporated into this ''one-dimensional vector'' while using the attention information. Such a global pooling plus 1*1 convolution structure can generate the channel-wise Attention mechanism and introduce global information. Two branches are added to the central part to weight the pre-fusion features and the fusion features to form the initial structure of AD-LinkNet.

C. MODULES OF AD-LINKNET

As shown in Figure 2, part A of AD-LinkNet is the Encoder of the network, which is based on pre-trained ResNet. ResNet itself is a neural network with particularly strong representation ability. Using ImageNet pre-trained ResNet as initialization can enhance the representation and generalization ability of AD-LinkNet, and can greatly improve the convergence speed of the network during training.

Part B is the central part of the network. This part uses the dilated convolution of the short-cut connection to form a series-parallel combination structure. And the channel-wise Attention mechanism is added before and after the dilated convolution.

The unfolded structure is shown in Figure 3. The central of part B is divided into five branches, each of which has a different depth and a different receptive field size. From top to bottom, the network's receptive field size for the input feature map is 31, 15, 7, 3, 1, and the depth is 4, 3, 2, 1, 0. If the depth is 0, it means that it is the identity map. This structure greatly enhances the receptive field of the central part of the network while maintaining the spatial resolution of the feature. At the same time, the features of different depths and different breadths are merged, so that the resulting feature map has sufficient receptive field and multi-dimensional semantic information. Finally, the feature scale is kept unchanged, and there is no loss of relative information in space.

We also add channel-wise before and after the central dilated convolution. Although channel-wise Attention mechanism can better synthesize all layers features, but this is not the main purpose of AD-LinkNet. AD-LinkNet's main purpose is to lead out branches from the network central part, so that the network can ''decouple'' the abstract information of the image, and the network can understand the meaning

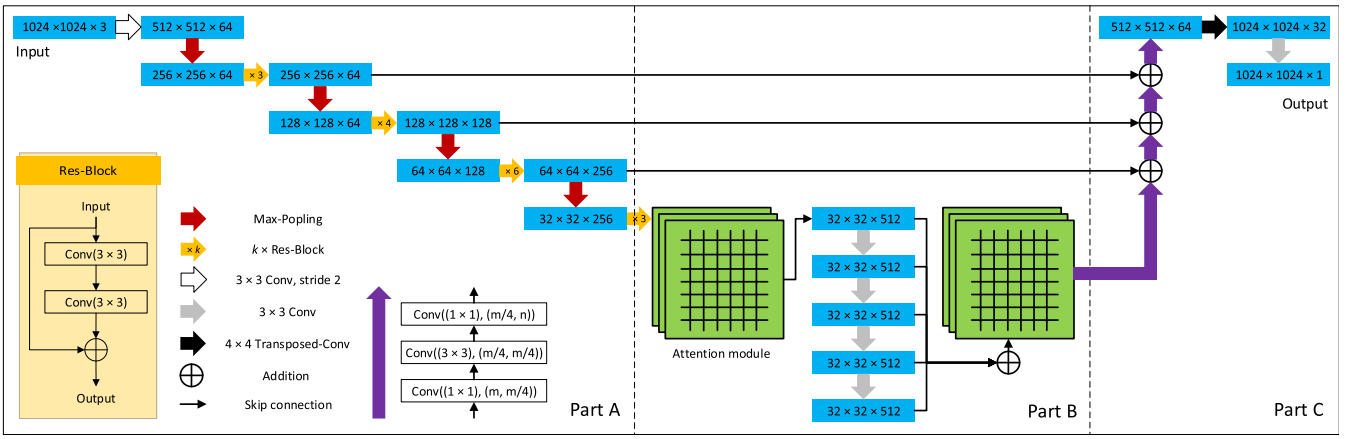


FIGURE 2. The structure diagram of AD-LinkNet.

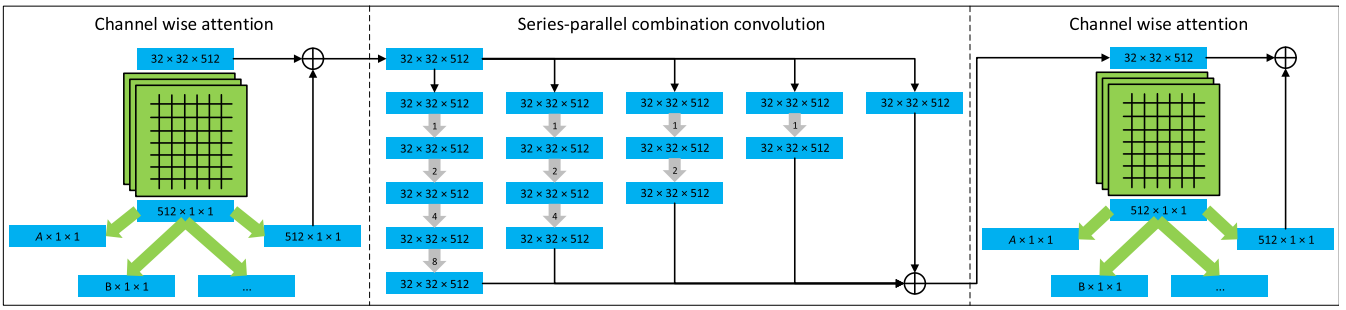


FIGURE 3. Schematic diagram of the AD-LinkNet central part.

of the image more comprehensively. In the central part of AD-LinkNet, several supervised branches are introduced to form a multi-task model.

As shown in Figure 3, AD-LinkNet adds channel-wise Attention mechanism and multi-task joint training before and after multi-layer feature aggregation. Different tasks can use different one-dimensional vectors lengths. Adding the module before and after feature aggregation allows different semantic layers of the network to have decoupled abstract logic information. This kind of branch structure is very suitable for weak supervised learning and semi-supervised learning. Weak supervised learning only needs to add weak labels as SE-Loss in training; unlabeled data can participate in training through GAN or Auto Encoder (AE). The branch of AD-LinkNet can be used as a discriminator for the GAN, or as a Decoder for the encoder.

Part C is the Decoder part of the network. This part remains consistent with LinkNet. The purple arrow part of Figure 2 uses the bottleneck structure of the residual network [36]. This structure reduces the overall computational load by introducing a 1*1 convolution kernel [37], and can increases the number of activation functions in the network and improves the network's representation ability. Part C uses transposed convolution for up-sampling, and up-sample the feature map by 32 times of the side length to restore the semantic label map with the same scale as the original image.

D. TRAINING

In this paper, three sets of satellite data are augmented using horizontal, vertical and diagonal folding methods. A total of 8 times of training data was obtained after augmenting compared to the original amount of data. The test image is deformed differently and then through the network, the output semantic map is restored to the original shape, and then they are combined. This method of augmenting does not take up training time, but it doubles the test time based on the augmenting data.

Since the satellite image has "isotropy", it has no so-called up, down, left and right points, and the data can be efficiently augmented by rotating the image. At the same time, since the satellite image is taken in a bird's-eye view state, most objects are stretched, and the semantics can be kept unchanged. Satellite images will have some light and dark changes due to different shooting time, and the land coverage of each location has a large difference. Therefore, we try to use a more radical color augmentation method, as shown in Figure 4. We change the hue of the original image from -30 to $+30$, the saturation of the original image from -5 to $+5$, and the value of the original image from -15 to $+30$. The first line of the figure is the change in hue, the second line is the change in saturation, and the third line is the change in value. We apply this color augmentation method to the training of multiple models in road extraction tasks and

FIGURE 4. Example of color augmentation of satellite imagery (the original image at the center).

land classification, and finally compare their test results. The method of color augmentation is verified to be universal for data augmentation of satellite images.

For the choice of loss function, the road occupy a small proportion of the overall picture, but the background have a large percentage in the road extraction dataset. For the land classification dataset, the proportions of the segmentation target and the background are also unbalanced. So we choose to use Dice loss instead of IoU loss.

IV. EVALUATION

In this section we first introduce three data sets and implementation details. Next, we performed an ablation study on the data augmentation and various methods used in the experiment. Based on the data augmentation and the specified loss function, we compare the different depth models of AD-LinkNet in many aspects. At the same time, we also compare the segmentation effects of AD-LinkNet and other models on the road extraction dataset, and verify the transfer learning ability and universality of AD-LinkNet on the land classification dataset.

A. DATASETS AND METRICS

This paper uses three satellite semantic segmentation datasets, namely DeepGlobe's road extraction dataset [1], DeepGlobe's land classification dataset [1], and Inner Mongolia's land classification dataset.

The DeepGlobe road extraction dataset is a 2-tiles dataset from Thailand, India, and Indonesia. The road extraction dataset contains more scenes and complex road conditions. The task of this data set is to extract hardened roads from satellite images. The data set contains 6226 pairs of training data, 1243 verified images, and 1101 test images. All image sizes are 1024*1024, and the ground resolution of the image pixels is 0.5m/pixel.

The DeepGlobe land classification dataset is a 7-tiles dataset that includes: urban land, agricultural land, pasture, woodland, water, barren land, and unknown land (including clouds and others) which contains 1146 satellite images with 2448*2448 pixels. It contains 803 pairs of training data, 171 verification images, and 1101 test images. All images contain RGB data with the ground resolution of the image pixels is 0.5m/pixel. Roads and bridges are not annotated in the training set because they are already reflected in the road extraction challenge.

The Inner Mongolia land classification dataset is a 7-tiles dataset from the Jilin No. 1 satellite. The original resolution of the satellite is 27338*24631 with total of 12 pictures, and the ground resolution of the image pixels is 0.7m/pixel. In this paper, the original image is reduced to 7k resolution test chart with resolution of 1024*1024. We use this dataset to test the trained model and visually compare the test results with the DeepGlobe land classification dataset to explore the impact of different satellite datasets on the model's performance.

For the Metric: In the road extraction task, we use the pixel-wise Intersection over Union (IoU) score as our evaluation metric for each image, defined as Eqn. (1) [1].

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{1}$$

And TPi is the number of pixels, which are correctly predicted as road pixels. FPi is the number of pixels, which are incorrectly predicted as road pixels. FNi is the number of non-road pixels, which incorrectly predicted as image i. We assume that there are n images, the final score is defined as the average IoU among all images (Eqn. (2)).

$$mIoU = \frac{1}{n}\sum_{i=1}^{n} IoU_i \tag{2}$$

In land classification task, we also use the pixel-wise Intersection over Union (IoU) score as our evaluation metric [1]. It was defined slightly differently for each class, as there are multiple categories (Eqn. 3). Assuming there are n images, the formulation is defined as,

$$IoU_j = \frac{\sum_{i=1}^{n} TP_{ij}}{\sum_{i=1}^{n}(TP_{ij} + FP_{ij} + FN_{ij})} \tag{3}$$

And TPij is the number of pixels in image i, which are correctly predicted as class j. FPij is the number of pixels in image i, which are incorrectly predicted as class j. FNij is the number of pixels in image i, which are incorrectly predicted as any class other than class j. Note that we have an unknown class that is not active in our evaluation. And the final score is defined as the average IoU among all classes as in Eqn. (2).

B. IMPLEMENTATION DETAILS

In the road extraction and land classification task. We use PyTorch as the deep learning framework. All models are trained on 4 NVIDIA GTX1080 GPUs. This paper mainly do

TABLE 1. Results on validation set of Ablation study in the DeepGlobe Road Extraction Task.

	Method (using road extraction data set)	IoU score (%)	Growing score (%)
0)	Deep Unet	58.29	—
1)	Deep Unet + TTA	61.76	3.47
2)	Deep Unet + TTA + Color Augmentation + IoULoss	62.15	3.86
3)	Deep Unet + TTA + Color Augmentation + InverseMask	62.52	4.23
4)	Deep Unet + TTA + Color Augmentation + Cancel Augmentation + Fine tuning	62.78	4.49
5)	Deep Unet + TTA + Color Augmentation + DiceLoss	62.94	4.65

the experiment of Deep Unet, LinkNet, D-LinkNet [8] and AD-LinkNet on the DeepGlobe road extraction and DeepGlobe land classification dataset. At the same time, different depths of AD-LinkNet are analyzed from various aspects such as network parameters, training efficiency and network accuracy. And the AD-LinkNet proposed in this paper is compared with our previously proposed D-LinkNet34.

In these experiments, the batchsize is set between 2 and 8 according to the network. In the transfer learning, we modified partial BatchNorm layer [38] of the network to the InstanceNorm [39]. The stability of the network can be improved when the input Batchsize is small. And using Adam and RMSProp as the optimizer, the initial learning rate is 1e-4, and the network loss function tends to be stable. The measurement standard is reduced to the one fifth of it. The baseline's (Deep-Unet) IoU score was 0.5829 without any data augmentation.

C. ABLATION STUDY

To study the effects of the individual data augmentation method and accuracy improvement method, this part shows an ablation study by systematically adding them one at a time. For this study, we chose Deep-Unet as the baseline model and the road extraction dataset as the experimental data set, the result is shown in Table 1 and Figure 5. At the same time, the results show that these methods are effective for improving the accuracy of semantic segmentation.

1) Augmentation at the time of testing (TTA). When TTA was added to Deep-Unet, the accuracy rate increased by 0.0347 relative to the baseline model. The IoU score is 0.6176.

2) Color Augmentation(including deformation augmentation). After adding TTA and color augmentation on Deep-Unet, IoULoss was used as a loss function. The accuracy rate increased by 0.0386 relative to the baseline model. The IoU score is 0.6215.

3) InverseMask. Retain TTA and color augmentation added on Deep-Unet. The values of 0 and 1 of the mask are reversed for model training, and the appropriate model fusion is attempted. The accuracy rate increased by 0.0423 relative to the baseline model. The IoU score is 0.6252.

4) Cancel augmentation for tuning. Retain TTA and color augmentation added on Deep-Unet. At the end of the training. The color augmentation is removed, retaining only the basic augmentation method like folding. But such an operation sacrifices the generalization ability of the network, allowing the network to converge more on the training set. The accuracy rate increased by 0.0449 relative to the baseline model. The IoU score is 0.6278.

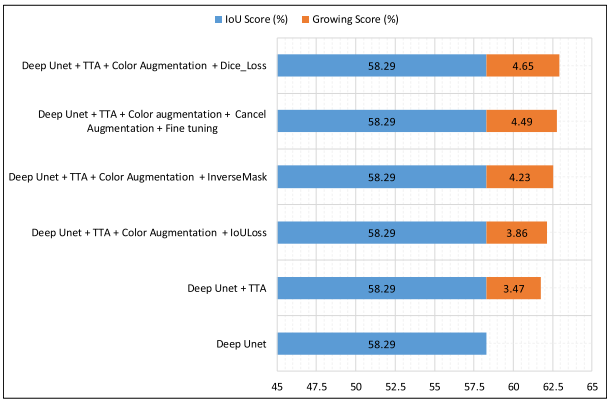


FIGURE 5. Ablation study of accuracy improvement method.

5) Replace the loss function. Retain TTA and color augmentation added on Deep-Unet. Replace the IoULoss loss function with the DiceLoss loss function. The accuracy rate increased by 0.0465 relative to the baseline model. The IoU score is 0.6249.

D. EXPERIMENTAL RESULTS

Through the superposition of the above methods, Deep-Unet has been able to achieve a fairly high segmentation accuracy, but it has some inherent problems. There are many misidentifications in the segmentation results of Deep-Unet, for example it recognizes rivers, fields and so on as roads. From the training loss of Deep-Unet, the final IoU score on the road extraction training set is 62.94%, which means that the network has reached a less ideal local optimum. Considering this problem, the network needs to be able to better grasp the inherent characteristics of the image during training. The other three networks in this experiment (LinkNet, D-LinkNet, and AD-LinkNet) use the transfer learning method to help the network to be converged. Under the premise of using all the effective method in section 4.3, and using the road extraction data set, the segmentation results of Deep-Unet, LinkNet, and D-LinkNet34 are shown in the Table 2. Meanwhile, in order to improve the performance of the competition, we tried different ways of model fusion.

As shown in Table 2, the segmentation result of LinkNet34 is equivalent to that of Deep-Unet. We quantify the segmentation results of the two and find that the

TABLE 2. Results on validation set of different models in the DeepGlobe Road Extraction Task.

Network structure (using road extraction data set)	IoU score (%)
Deep Unet	62.94
LinkNet34 (pre-training)	63.00
D-LinkNet34 (pre-training)	64.12
Model Ensemble of Deep Unet and LinkNet34 (pre-training)	63.94
Model Ensemble of Deep Unet, LinkNet34 (pre-training) and D-LinkNet34 (pre-training)	64.49

mIoU value of the segmentation results is 0.785, which means that there are large differences between the two networks. The advantage of Deep-Unet is that its larger receptive field can grasp more global information. The advantage of LinkNet is that the feature resolution of the network center is relatively high, and the Encoder part of the network uses pre-trained ResNet34, which has stronger recognition ability. For our previous proposed D-LinkNet34, the accuracy of segmentation result exceeds the segmentation accuracy of the original model. As shown in last two line of the table, the model fusion method is effective for improving the segmentation accuracy in the competition.

This paper attempts to use multiple depths of ResNet as AD-LinkNet different pre-training Encoder and different Decoders, this paper tests various depths of AD-LinkNet on road extraction datasets, including AD-LinkNet34, AD-LinkNet50 and AD-LinkNet101. We comprehensively analyze the performance of each network from the perspectives of accuracy, parameter quantity, and training time.

As can be seen from Table 3, AD-LinkNet34 is a relatively comprehensive network in terms of overall parameter quantity, training time and accuracy. The most accurate network is the AD-LinkNet50, but the parameter size is 7 times that of the D-LinkNet34, the training time is about 4 times, and the forward prediction time is about 4 times that of the AD-LinkNet34. AD-LinkNet101 has the highest depth and maximum parameter, but its performance is not better than AD-LinkNet34. This is because the data of the 6226 training data of this data set is not enough to make AD-LinkNet101 converge completely. In summary, for the scenario where the number of split semantic labels is limited, we choose AD-LinkNet50 with shallow depth and highest precision as the final competition model, and we got the 1st place in the DeepGlobe road extraction challenge.

In order to verify the performance of the AD-LinkNet network. We use the same data augmentation and loss function to compare the accuracy of AD-LinkNet and our previously proposed D-LinkNet34, and the two DeepGlobe data sets of road extraction and land classification are used to verify the ability of transfer learning and the universality of practical application.

As shown in Figure 6, AD-LinkNet improve 0.61% accurate than D-LinkNet34 in road extraction tasks. At the same time, AD-LinkNet get 0.35% higher than D-LinkNet34 in the land classification task. It can also be seen that the AD-LinkNet have better transfer learning ability and application universality for satellite semantic segmentation tasks.

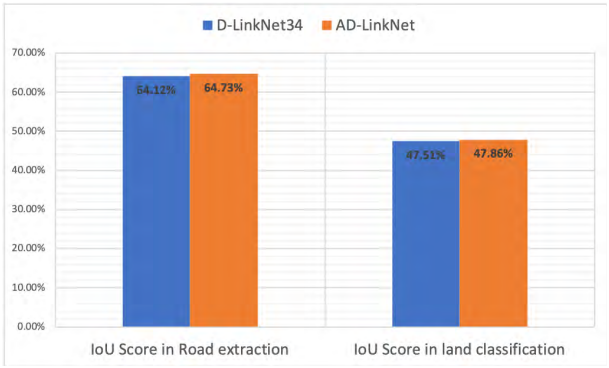


FIGURE 6. Segmentation performance of D-LinkNet and AD-LinkNet on road extraction and land classification datasets.

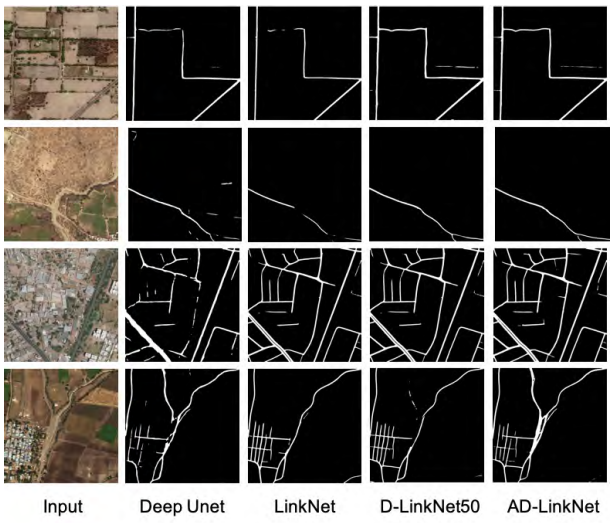


FIGURE 7. Test results of different models on road extraction tasks.

E. RESULT ANALYSIS

For the task of road extraction, as shown in Figure 7, the first two lines of the figure show the road connection problem in LinkNet, and there are several road interruptions in the segmentation result of LinkNet, while there is no such problem in Deep-Unet, D-LinkNet, and AD-LinkNet. The last two lines are examples of Deep-Unet mispredictions. Deep-Unet is more likely to mistake the road as a background or treat a non-road like river as a road (the third line and fourth line, many buildings between roads are not identified). D-LinkNet50 and AD-LinkNet not only have Deep-Unet's large receptive field, but also have LinkNet's pre-trained Encoder and high-resolution center feature map, and its

TABLE 3. Results on validation set of different depth AD-LinkNet in the DeepGlobe Road Extraction Task.

Network structure(Road extraction)	IoU Score (%)	Parameter quantity	Training time (GPU*H)
AD-LinkNet34	64.73	119M	2*35H
AD-LinkNet50	64.79	831M	4*70H
AD-LinkNet101	63.37	904M	2*120H

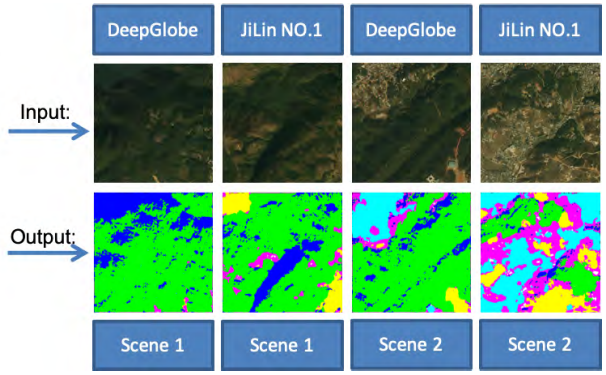


FIGURE 8. D-LinkBrach test results for similar terrain on different satellite land classification datasets.

unique multi-scale feature fusion, thus avoiding disadvantages of Deep-Unet and LinkNet and made a better prediction. Compared to D-LinkNet50, AD-LinkNet is more precise in handling small routes and can accurately segment branch routes along the main road (as shown in the fourth line).

For land classification tasks, it is difficult to accurately segment the forest along with the terrain of the lake (scene1). Conversely, the terrain associated with the forest and the city is easier to segment (scene2). Therefore, we select two kinds of terrain distribution images, and use AD-LinkNet to test the DeepGlobe land classification dataset and the Inner Mongolia land classification dataset. The test result chart is shown in Figure 8. The data of two different data sets are derived from two different satellites. The biggest difference between the two sets of data is the different original resolution, such as the ground resolution of the image pixels, and the color and brightness of the picture. It is not difficult to find that the pixel resolution (0.5m/pixel) of the DeepGlobe dataset is smaller than the pixel resolution (0.7m/pixel) of the Inner Mongolia dataset. At the same time, the Color of the DeepGlobe dataset is relatively softer and brighter. For first two columns in Figure 8, AD-LinkNet can make a clear segmentation on DeepGlobe for the terrain of the forest with the lake. However, in the Inner Mongolia dataset, it mistakenly predicts the forest under the shadow as a lake (actually there is no lake in the image). For last two columns in Figure 8, AD-LinkNet can segment forests and cities on both the DeepGlobe dataset and the Inner Mongolia dataset, but the edge processing on the DeepGlobe dataset is more accurate. As shown in the fourth image, It mistakenly predicts the edge between urban and forest as rangeland. So we think that the pixel resolution of the dataset and the color and brightness of the image have an impact on the model segmentation. For different data sets, different model optimizations and fine tune should be done.

V. CONCLUSION

In this paper, we focus on the refinement of satellite image semantic segmentation. Through network design and loss function design, the segmentation result is more precise and detailed. Another work in this paper is to design a data processing and transfer learning method to reduce the semantic label requirements of the image semantic segmentation task in the satellite domain. In terms of data processing, we design the universal data augmentation method of image morphology, color augmentation, and TTA. For refined semantic segmentation, we use LinkNet as the basis model and use pre-trained ResNet as Encoder to implement transfer learning. We designed a combination module (AD-Link), which includes a series-parallel combination dilated convolution and two channel-wise Attention mechanism, and add AD-Link to the central part of AD-LinkNet. Meanwhile, based on road extraction and land classification satellite image, we conducted experiments on two representative satellite domain tasks. Subsequently, we compared various networks to clarify the importance of the receptive field and the feature map resolution, and verified the validity of the AD-Link structure and the AD-LinkNet network.

The various satellite image semantic segmentation networks described in this paper are fully convolutional structures, most of which do not contain a global pooling structure in the central part. For a network without global pooling, the process from the original image to the semantic image is a fixed-size image. The mapping to a pixel label is similar to patch-based segmentation. The learning process of the network is still a fitting of the data itself, but the full convolution structure can realize the weight sharing in the calculation. The information with global pooling structure and coupled non-fixed scale must change the mapping mode from patch to pixel. The similar information coupling method has been applied to the object detection field [40]. Then we will explore and research a variety of information coupling methods. Finally, our future research direction will also involve multiple directions of image processing [41]–[43].

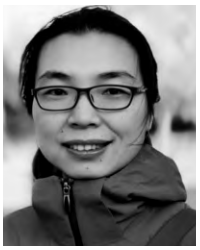
REFERENCES

[1] I. Demir *et al.*, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, May 2018, pp. 172–209.

[2] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.

[3] T.-Y. Lin *et al.*, "ar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2014, pp. 740–755.

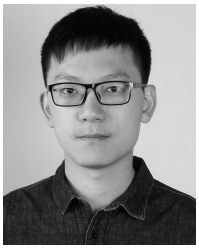
[4] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 180–196.
[5] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 770–778.
[7] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223.
[8] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 182–186.
[9] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122
[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.
[11] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2017, pp. 5659–5667.
[12] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[14] V. Iglovikov and A. Shvets. (2018). "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation." [Online]. Available: https://arxiv.org/abs/1801.05746
[15] L. Barghout and L. Lee, "Perceptual information processing system," U.S. Patent 10 618 543, Mar. 25 2004.
[16] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.
[17] A. Huertas and R. Nevatia, "Detecting buildings in aerial images," *Comput. Vis., Graph., Image Process.*, vol. 41, no. 2, pp. 131–152, Feb. 1988.
[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[19] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2016, pp. 3640–3649.
[20] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2017, pp. 11–19.
[21] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1568–1576.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*. New York, NY, USA: Springer, 2015, pp. 234–241.
[23] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.
[24] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2016.
[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "Overfeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229
[26] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 472–480.
[27] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.
[28] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[29] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.
[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.
[32] S. Zagoruyko and N. Komodakis. (2016). "Wide residual networks." [Online]. Available: https://arxiv.org/abs/1605.07146
[33] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 550–558.
[34] Z. Wu *et al.*, "Blockdrop: Dynamic inference paths in residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8817–8826.
[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[36] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.
[37] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400
[38] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[39] D. Ulyanov, A. Vedaldi, and V. Lempitsky. (2016). "Instance normalization: The missing ingredient for fast stylization." [Online]. Available: https://arxiv.org/abs/1607.08022
[40] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 734–750.
[41] Q. Hu, H. Wang, L. Teng, and C. Shen, "Deep cnns with spatially weighted pooling for fine-grained car recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3147–3156, Nov. 2017.
[42] Z. Ma, Y. Lai, W. B. Kleijn, Y. Z. Song, L. Wang, and J. Guo, "Variational bayesian learning for dirichlet process mixture of inverted dirichlet distributions in non-gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2018.
[43] Z. Ma, J. H. Xue, A. Leijon, Z. H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2016.



MING WU was born in Shanghai, China, in 1977. She received the M.S. and Ph.D. degrees in information and communication engineering from the Beijing University of Posts and Communications, Beijing, China, in 2003 and 2012, respectively, where she is currently an Associate Professor with the School of Information and Communication Engineering. She has published over 20 technical papers in international journals and conferences. Her primary research interests include computer vision, pattern recognition, deep learning, satellite image parsing, and human gesture recognition.

CHUANG ZHANG was born in Changchun, China, in 1975. He received the B.S. degree in mechanical engineering from the Jilin University of Technology, Changchun, in 1998, the M.S. degree in mechanical engineering from Jilin University, in 2001, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), in 2004, where he was an Associate Professor with the Laboratory of Pattern Recognition and Intelligence System (PRIS), School of Information and Telecommunications Engineering, since 2006. He has published over 50 technical papers in international journals and conferences. His research interests include computer vision and pattern recognition, satellite image parsing, and human gesture recognition.

JIAMING LIU was born in Beijing, China, in 1996. He has studied e-commerce and law at the Beijing University of Posts and Telecommunications and also has studied this major at the Queen Mary University of London. He received diplomas and degree certificates from two schools. In 2018, he represented BUPT to conduct research exchange at the Robot Laboratory, Queen Mary University of London. He is currently pursuing the master's degree in information and communication technology with the Beijing University of Posts and Telecommunications. He is also a member of the Pattern Recognition and Intelligent System Laboratory. His research projects include robotic visual positioning, and tracking and robotic autonomous movement display. His current research interests include object detection, semantic segmentation, and data augmentation.



LICHEN ZHOU was born in Hunan, China, in 1994. He is currently pursuing the master's degree in information and communication engineering with the Beijing University of Posts and Telecommunications. His research interests include computer vision, deep learning, image parsing, and face recognition.

XIAOQI LI is currently pursuing the bachelor's degree in information engineering with the Beijing University of Posts and Telecommunications. Her current research interests include bioinformatics and computer vision, especially medical instruments that are relevant with vision.