Road Extraction From High Spatial Resolution Remote Sensing Image Based on Multi-Task Key Point Constraints

XUNGEN LI\textsuperscript{1,2}, ZHAN ZHANG\textsuperscript{1}, SHUAISHUAI LV\textsuperscript{1,2}, MIAN PAN\textsuperscript{1}, QI MA\textsuperscript{1}, AND HAIBIN YU\textsuperscript{1}

\textsuperscript{1}College of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China
\textsuperscript{2}Pujiang Microelectronics and Intelligent Manufacturing Research Institute, Hangzhou Dianzi University, Jinhua 322200, China

Corresponding author: Shuaishuai Lv (lvshuai@hdu.edu.cn)

This work was supported in part by the National Key Research and Development Project of China under Grant 2016YFC1400302, in part by the National Natural Science Foundation of China under Grant 61501155 and Grant 61871164, and in part by the National Defense Science and Technology Key Laboratory Fund under Grant 6142401200201.

\textbf{ABSTRACT} To solve some problems of high spatial resolution remote sensing images caused by land coverage, building coverage and shading of trees, such as difficult road extraction and low precision, a road extraction method based on multi-task key point constraints is put forward in this article based on Linknet. At the preprocessing stage, an auxiliary constraint task is designed to solve the connectivity problem caused by shading during road extraction from remote sensing images. At the encoding & decoding stage, first, a position attention (PA) mechanism module and channel attention (CA) mechanism module are applied to realize the effective fusion of semantic information in the context during road extraction. Second, a multi-branch cascade dilated spatial pyramid (CDSP) is established with dilated convolution, by which the problem of loss of partial information during information extraction from remote sensing road image is solved and the detection accuracy is further improved. The method put forward in this article is verified through the experiment with public datasets and private datasets, revealing that the proposed method provides better performance than several state of the art techniques in terms of detection accuracy, recall, precision, and F1-score.

\textbf{INDEX TERMS} Road extraction, remote sensing image, semantic segmentation, high-resolution imagery.

I. INTRODUCTION

In modern society, roads are an important identification object in maps and geographic information systems. As a kind of important infrastructure, roads play an important role in the construction of cities and towns, traffic and transportation and other fields [1]–[4]. So, how to extract road information from remote sensing images quickly and accurately has aroused the attention of many scholars [5]–[8].

During road extraction from remote sensing images, the preprocessing of data is usually carried out first to extract road features, and roads are classified based on pixel points to get the final result. According to different research methods of road extraction, road extraction methods can be classified into the traditional method and the deep learning method.

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni.

With the traditional method, roads are usually recognized and extracted by establishing a feature model based on the basic feature of roads and human experience. Common features include the background feature [9], path morphology [10], feature template [11] and dynamic outline feature [12]. Chaudhuri et al. [13] extracted roads by enhancing morphological features of directions and deduced road sections accurately based on the feature of road sections. Miao et al. [14] put forward a semi-automatic road detection method based on mean shift where the initial road section of the preset road seed point was extracted with the geodesic method, and then roads and non-roads were segmented with the threshold value to extract the smooth and correct central line of roads. Unsalan and Sirmacek [15] extracted the initial edge feature of roads and then extracted roads with the binary Ballon algorithm and the graph theory, reducing time spent on road extraction. Song and Civco [9] put forward the method of detecting road regions with regional growing technology...
based on support vector machine [10] and similarity criterion respectively. With this method, road features can be extracted through simple threshold segmentation of the shape index and the density feature of roads, achieving good effects in road extraction from images composed of the urban and rural features.

The traditional method has played an active role in certain application scenarios, but many threshold parameters involved in feature design and connection algorithm can achieve good effects in some types of images only by artificial adjustment, which restricts its application to large-scale data.

Compared with the traditional method, the road extraction method based on deep learning has huge strengths. Convolutional networks in deep learning can explore high-level features of roads to improve the effectiveness of computer vision tasks. And it also has a powerful self-learning ability and an efficient feature fitting ability, showing huge strengths in road extraction precision and automation. For example, Sun et al. [16] proposed a road extraction method based on stacked U-nets and a multi-output network model, solving the imbalance problem of samples in multi-goal classification. Chen et al. carried out super-resolution accurate extraction of narrow roads with one-class support vector machine (OCSVM) classifier, by which the road extraction could only extract roads whose width was only several pixels. Mnih and Hinton [17] proposed a method where road regions in high-resolution aerial imagers were detected with constrained Boltzmann machine, and the number of dimensions of input data was reduced by preprocessing to make segmented lines smoother. Saito et al. [18] designed a model for road recognition based on extraction of feature of sliding windows, refining the extraction process of road extraction.

However, the above methods based on deep learning are only mechanical application models which do not consider the scenario of road extraction, the resolution of feature maps will decline gradually during down sampling, and then information of roads whose width contains fewer pixels will be lost severely. Some lost road information cannot be recovered during up sampling effectively, which will significantly reduce the performance of road extraction in networks.

As deep learning develops rapidly [19]–[21], full convolutional networks (FCN), such as Alexnet [22], VGGNet [23] and ResNet [24], have shown outstanding feature information extraction ability which promotes further development of the road extraction technology. Zhou et al. [25] fused ResNet with DLInknet structure, put forward by LinkNet [26], expanded the convolutional layer in the central part, and widening the receptive field [27] of feature points. Li et al. [28] improved the network model in [25] and established a D-Linknetplus model based on DBlockPlus which reduced model parameters and improved the calculation efficiency. The above two methods only pay attention to local road extraction and do not take into consideration the distinctiveness of road scenarios, so roads in road images extracted will be not continuous. Kestur et al. [29] designed a model of road extraction from remote sensing images based on the U-shaped full convolutional network (U-FCN) model. This model is established based on the classical FCN [30] network and is composed of a group of stacked convolution and corresponding stacked mirroring deconvolutions [31]. This method merely uses a general segmentation network structure to finish road extraction tasks, but wastes a large quantity of semantic information on remote sensing road images. Hu et al. [32] designed a graph generation model based on deep learning and recurrent attention network, which could learn by itself the road generation mode according to the probability distribution of road edges in road images to get the final generated road map. However, both of them do not consider the role of road intersection during road extraction, which can be fused to improve the connectivity of road extraction.

Although the above method has made great progress in terms of road extraction from remote sensing images, there are still many problems during road extraction from remote sensing images, such as building coverage, shading of trees and diversity of roads, bringing many challenges to extraction and segmentation of road details. In essence, road extraction is binarization segmentation of roads in remote sensing images, which is one of the semantic segmentation tasks. Therefore, there are many new semantic segmentation models in the field of semantic segmentation. Yu et al. proposed the BisNet [33] model and the DFN [34] model to achieve the image semantic segmentation respectively. The BisNet completed the extraction of image features by constructing a feature fusion module and a refinement module. And the discriminative feature network (DFN) constructed two sub-networks to solve the problem of intra-class inconsistency and inter-class indistinction. Zhang et al. [35] built an EncNet model, which introduced a context encoding module to capture the context information of the image. Zhang et al. [36] and Di et al. [37], respectively constructed an EncNet model and segmentation framework MSCI to improve image segmentation performance. However, the above segmentation models are usually extracted local target objects, which cannot effectively address the difficulties of road extraction. The difficult point of road segmentation is that the road structure is complicated. First, roads are usually similar to the surrounding environments (for example, grains of building roofs are usually similar to those of roads in city images; Traces left by wind are similar to those of roads in desert image). Secondly, the roads may be blocked by barriers, such as shadows or visual shading, which will be recognized roughly (for example, shadows of buildings will block roads in city images), and recognition of roads et.al, which cannot be seen is particularly difficult, for example, the roads with good afforestation will be covered by the trees, causing the road to becoming completely invisible.

Linknet is an effective semantic segmentation neural network, which uses the advantages of skip connections, residual blocks, and codec structure. Linknet also has fewer
The structure of the article is as follows. Part II briefly reviews the basic framework of this article; Part III introduces the general methodology of methods mentioned herein and the road extraction procedure; Part IV displays the experiment result of methods and the result of the comparison experiment; Part V is the summary of the whole article and outlook on future work.

II. BRIEF REVIEW

Linknet CNN is used as the basic network of road extraction in this article. The structure is as shown in Figure 1. The main contributions of this article are as follows:

I. A multi-task intersection point constraint (IPC) module is designed. The road extraction ability is improved in complicated scenarios by adding an auxiliary constraint task into the segmentation task and calculating the semantic constraint angle between road points.

II. The position and channel attention (PCA) modules are considered to improve the road extraction precision in high-resolution images, which fused semantic information in the context during semantic segmentation and filtered other redundant information of remote sensing road images.

III. A multi-branch CDSP module which contains four different branches is established. As regards each branch, feature extraction in different scales is carried out for the feature map, feature extraction results of each of the four branches are fused to take into consideration diversified level features in the model and recover original feature of images at the decoding stage, especially for the feature of roads with a small width in images.

The main contributions of this article are as follows:

I. A multi-task intersection point constraint (IPC) module is designed. The road extraction ability is improved in complicated scenarios by adding an auxiliary constraint task into the segmentation task and calculating the semantic constraint angle between road points.

II. The position and channel attention (PCA) modules are considered to improve the road extraction precision in high-resolution images, which fused semantic information in the context during semantic segmentation and filtered other redundant information of remote sensing road images.

III. A multi-branch CDSP module which contains four different branches is established. As regards each branch, feature extraction in different scales is carried out for the feature map, feature extraction results of each of the four branches are fused to enhance the ability to recover original information of images during up-sampling.

The structure of the article is as follows. Part II briefly introduces the basic framework of this article; Part III introduces the general framework of methods mentioned herein and the road extraction procedure; Part IV displays the experiment result of methods and the result of the comparison experiment; Part V is the summary of the whole article and outlook on future work.

III. METHODOLOGY

The road extraction procedure is as follows: First, data enhancement means are applied to input high-resolution remote sensing road images, such as HSV contrast ratio transformation, geometric transformation, image cropping, etc. Second, the feature of input images is extracted with Encoder, and the feature map obtained will pass the PCA module where a large quantity of semantic information of roads in different positions and different channels can be fused effectively. Then, the feature map will pass the CDSP
module where the problem of severe loss of road information during down-sampling of CNN. At last, the feature map passes two up-sampling feature recovery decoders with a similar structure, i.e. IPC Decoder and Seg Decoder. As the main task, Seg Decoder can output the pattern segmentation result, while IPC Decoder, as the auxiliary task, can output the matrix of semantic constraint angle results of images.

The input of the model has two parts: the label of the original image and each pixel. Under such a framework, the road image extraction can be converted into loss calculation of labels of images extracted by CNN based on original images. Such labels are corresponding to spatial dimensions $w \times h$ of original images one by one. $w$ and $h$ represent the width and height of the image. In the label images, 0 and 1 represent the background and the road to be extracted respectively.

In terms of feature extraction, CNN performs well in terms of the segmentation task, such as Unet [41], Densenet [42], and Depplabv3+ [39]. So, the encoder feature extraction module and CNN mentioned are selected for extracting the road image feature in this article.

Next, we will elaborate on the composition of the network module of each part in the framework. First, the establishment mode of the multi-task constraint module is described. Second, the specific establishment process of PCA is discussed. At last, the composition principle of multi-branch CDSP is described in detail.

### A. MULTI-TASK IPC

Figure 2 reveals that IPC output is a feature matrix of semantic constraint angle. The IPC model is an additional task related to prior information in order to solve the problem of road disconnection in road extraction. As shown in Figure 3, the general network model is likely to cause road extraction disconnection because the road is covered by a tree. From Figure 3(c), it can be seen that the role of the auxiliary task constraint module is to provide the prior features of the road by calculating the semantic constraint angle. This constraint task can also allow the network to learn this angle autonomously by calculating the angle between each pixel in the road marking map and its neighboring road. And then some deep-level feature information is generated, and this feature information cannot be perceived by our human intuition. The auxiliary task of calculating the semantic constraint angle solves the problem of road disconnection of remote sensing road image extraction. This problem of road disconnection is solved effectively.

Key pixel points of roads (i.e. road intersection points) in the prediction matrix figure and the ground truth semantic constraint angle matrix figure are constrained via the loss function. First, the skeleton of label images of roads is extracted. The skeleton map of roads is extracted with Zhang and Suen [43] burning algorithm. Each iteration step of the algorithm is corrosion of target pixels which conform to specific conditions to detail goals, i.e. road pixel labels in this article. Iterations are carried out continuously until goals (i.e. pixel points) corroded last time are not corroded in this round of operation. And then the ground truth semantic constraint angle is calculated. The feature map of pixel points is as shown in Figure 4. This algorithm should meet the four conditions of Formula (1), Formula (2), Formula (3) and...
intersection pixel or edge pixel is backtracked. If intersections are found in neighbor domains, the traversed edge serves as the connection edge of these two intersections.

After that, the semantic constraint angle of each road point is calculated: first, the road intersection extracted on the skeleton map of each connected road is marked as \([z_1, z_2, \ldots, z_k]\), pixel points between two neighbor road intersections extracted are marked as \([p_1, p_2, \ldots, p_n]\), the shortest path distance and azimuth angle of two intersections nearby each pixel point \(p_i\) are calculated respectively, and the azimuth angle corresponding to the intersection with the smallest shortest path distance is taken as the semantic constraint angle of this point.

Here, the A* [44] algorithm is used to calculate the shortest path distance. A* algorithm will find the shortest path from the starting point X to the ending point Y. The path calculation formula is as follows:

\[
F(n) = G(n) + H(n)
\]  

(5)

where \(G(n)\) represents the shortest distance from the starting point to the current point (straight line or diagonal line available), \(H(n)\) represents the Manhattan distance from the current point to the endpoint [45]. Only the horizontal distance and the longitudinal distance are calculated, other oblique distances are not calculated, and possible barriers are not taken into consideration.

The calculation process is as follows:

1. NewList (for storing coordinates of pixel points which may be involved), OldList (for storing coordinates of pixel points which can be not taken into consideration during calculation) and ShortList (for storing the shortest node distance between two points) are established. Set the current node as \(Z_i\).

2. Let current node \(Z = X\), and judge in sequence whether eight neighbor points \(Z_i(1 \leq i \leq 8)\) of \(Z\) can be accessed (that is, whether they are road pixel points). If \(Z_i\) cannot be accessed, \(Z_i\) is stored in the OldList (if \(Z_i\) has been in the OldList, \(Z_i\) is ignored). If \(Z_i\) can be accessed and is in the NewList, the distance between \(Z_i\) and \(Z\) is calculated and marked as \(g_i\), \(\text{New}_G(Z_i) = G(Z_i) + g_i\) and \(\text{New}_F(Z_i) = \text{New}_G(Z_i) + H(Z_i)\) are also calculated. If the new \(F\) value is greater than the old \(F\), the new \(F\) is adopted to be the new \(\text{New}_F(Z_i)\), and \(Z\) is set as the father node of \(Z_i\). If \(Z_i\) can be accessed, but is not in the NewList, \(Z\) is set as the father node and \(Z_i\) is added into the NewList.

3. Repeat steps (1)-(2), until the calculation of all the eight neighbor points is finished. Add \(Z\) into the OldList, traverse all the points in the NewList and select the minimum \(F\) value, mark this \(F\) value as \(\text{Min}_F\) and take it as the new \(Z\) node and then repeat the above steps.

4. When \(Z = Y\) node, the search is finished, and recursion starts to be conducted to search the father node of \(Y\) node.

5. Set the current node set as \(T\). Initialize \(T\) to be \(Y\), i.e. \(T = Y\). Add every node represented by \(T\) node into the ShortList and then make \(T\) be equal to the father node of the current node, i.e. \(T = T\).father.
the definition of azimuth angle and calculation method and the point on the road is the intersection obtained with the shortest path distance constraint angle. Assuming that the coordinate of \( \phi \) via the above calculation method and select the smaller \( \phi \) from \( \phi_i \) and its two neighbor intersections, i.e., the PCA module in Figure 2. This module includes the penetration point of improving the precision of road extraction.

On this basis, PCA is introduced into the model in this article, i.e., the PCA module in Figure 2. This module includes the CA mechanism and the PA mechanism. The attention mechanism module can make the network model pay attention to features and functions of such semantic information of roads on the level of positions and channels more effectively during road feature extraction, and the effect of PCA will be more significant, especially when there is more semantic information in remote sensing images.

The position attention module is as shown in Figure 5. Features of any position are updated via weighted aggregation of road features of all the positions. The weight needed in the update is determined by the similarity of the feature of roads at two positions. As they are more similar to each other, the weight will be greater; as they are less similar to each other, the weight will be smaller. The weight is irrelevant to the distance of two-point pixels.

Figure 6 shows that when one road feature map \( X \) is input, \( X \in \mathbb{R}^{C \times H \times W} \). Wherein, \( C \) represents the number of channels, \( H \) represents the height of the image, \( W \) represents the width of the image. Three new feature maps \( A \), \( B \) and \( D \) are obtained through three convolutions \[47\] with BN layer \[46\] and ReLU layer respectively. \( A \) and \( B \) are reshaped respectively to get \( \mathbb{R}^{C \times N} \). The definition of PA mechanism is as follows:

\[
S_{ij} = \frac{\exp(A_{ij}^T \cdot B_{ij})}{\sum_{i=1}^{N} \exp(A_{ij}^T \cdot B_{ij})}
\]

where \( S_{ij} \) represents the impacts of the \( i \) position on the \( j \) position. As road features at these two positions are more similar to each other, this value will be impacted more. Road feature map \( D \) is reshaped to get \( \mathbb{R}^{C \times N} \) and \( D^* \). The calculation formula of \( Y \) is as follows:

\[
Y_j = \alpha \sum_{i=1}^{N} \left( S_{ij}^T \cdot D_i^* \right) + X_j
\]

where \( \alpha \) is initialized to be 0 and is obtained through learning. Formula (10) shows that the final feature \( Y \) is updated via weighted aggregation of road features at all the positions in the image, so it can aggregate global semantic information.

The structure of the CA mechanism is shown in Figure 7. The CA mechanism is similar to the PA mechanism. They just emphasize different perspectives. For any two channels,
as long as their road feature is similar, a higher weight can be obtained. Each channel can be seen as a semantic response to different road features. The weight of channels can represent the dependence degree of semantic responses of specific roads. As the weight is greater, they are more dependent on each other; as the weight is smaller, they are less dependent on each other. Via the CA mechanism, the network can learn by itself the ability to express the semantic feature of a specific road.

The definition of the CA mechanism is as follows:

$$H_{ji} = \exp(A^T_i \cdot B_j)$$

$$Y_j = \beta \sum_{i=1}^{N} (H_{ji}^T \cdot D_i) + X_j$$

where $H_{ji}$ represents the impacts of the $i$ channel on the $j$ channel. The final result $Y_j$ is:

where $\beta$ is initialized to be 0 and is updated via training. It can be seen that the final feature is updated via weighted aggregation of road features on all the channels, so the final feature can aggregate global semantic information on the channel feature map.

At last, the feature map of PA mechanism and that of CA mechanism are fused. The above two feature maps are stacked with the dimensional stacking method to ensure that the original feature will not be damaged and retained completely and that networks can learn by itself how to integrate such complete road feature information during interaction.

### C. CDSP

The CDSP module is after the PCA module, as shown in Figure 2. To solve the problem of loss of local spatial information of roads caused by pooling operation and the problem that detailed features of roads included in images cannot be restored completely during up-sampling, a multi-branch CDSP is designed in this paper. The CDSP module uses hollow convolution to solve the problem of the traditional convolutional network in the feature extraction, which leads to the continuous decline of image resolution and the loss of spatial abstract information. The structure of CDSP is as shown in Figure 8. This module includes four branches: dimensionality reduction branch, self-adaptive pooling branch, even number of cascade branch of dilated convolution, and an odd number of cascade branch of dilated convolution.

1. Dimensionality reduction branch: The first branch is the dimensionality reduction branch of convolution. This branch is constructed with standard $1 \times 1$ convolution. It is mainly used to reduce the dimensionality of feature maps. This branch is equal to feature information of the original
road feature map. After finishing convolution, a BN layer and a ReLU layer are included.

(2) Self-adaptive pooling branch: A self-adaptive global pooling branch with convolution kernel as large as its size is constructed. The definition of convolution calculation is as follows:

$$\text{output} = \frac{n - k + 2 \times p}{s} + 1$$  \hspace{1cm} (13)

where $n$ is the size of the input feature map, $k$ is the size of the convolution kernel, $p$ is the filling range, and $s$ is the step length of convolution. Since there is one global pooling branch, $k$ is the size of the input feature map, $p$ is defaulted as 0, and $s$ is defaulted as 1. The convolution kernel is as large as its own size, so one time of up-sampling should be conducted before channel stacking with other branches.

(3) Even number of cascade branches of dilated convolution. Three dilated convolution modules with different dilated ratios are established respectively to improve the size of the receptive field of the output feature map. The dilated ratio is 2, 4, and 8 respectively. Three convolution modules with different dilated ratios are cascaded. The output definition of the dilated convolution is as follows:

$$\text{output} = \frac{n - (k - 1) + 2 \times p - 1 - d}{s} + 1$$  \hspace{1cm} (14)

where $d$ represents the dilated ratio. Definitions of other parameters are the same as those in Formula (14). To maintain that the size of the output feature map is the same, when the convolution kernel is 3*3, $p = d$, and $s = 1$, the size of the output feature map remains unchanged. When using the dilated convolution, the calculation definition of receptive field with dilation is as follows:

$$F(i, j - 1) = (F(i, j) - 1) \times s + d \times (k - 1) + 1$$  \hspace{1cm} (15)

where $F(i, j)$ represents the local receptive field of the $j$ layer under the impact of the $i$ layer, $k$ represents the size of the convolution kernel which is set as 3 in this article, $s$ represents the step length of the convolution, and $d$ represents the dilated ratio. It can be seen from the above formula that as $d$ becomes greater, the receptive field will be larger.

(4) The odd number of cascade branches of dilated convolution is similar to the third branch. The difference lies in those three dilated ratios are 3, 5, and 9 respectively. Other operations are similar to the third branch.

As regards different pyramid branches, different sub-domains are segmented for feature mapping, enabling the output of different sizes of feature mapping on different levels and at different positions in the pyramid pooling module. To maintain the weight of the global feature, if the pyramid has $N$ levels, $1 \times 1$ convolution is used after each level, the number of channels of the corresponding level is reduced to $1/N$, and then up-sampling is conducted for low-dimensional feature maps via bilinear interpolation to get the feature map with the same dimensions as the original feature mapping. At last, different levels of feature dimensions are stacked as the final global output of the pyramid pooling module.

D. LOSS FUNCTION

Dice Coefficient Loss+BCE Loss is used as the loss function of the main tasks of the model in this article. Its definition is as follows:

$$\text{DCLoss} = 1 - \frac{2 \sum_{i=1}^{N} |P_i \cap G_i|}{\sum_{i=1}^{N} |P_i| + |G_i|}$$ \hspace{1cm} (16)

$$\text{Loss} = \text{DCLoss} + \sum_{i=1}^{N} \text{BCELoss} (P_i, G_T)$$ \hspace{1cm} (17)

where $P_i$ represents the $i$ predicted image, $G_i$ represents the $i$-label image, and $N$ represents the Batch Size. The numerator represents the total number of correct predictions of the positive sample, and the denominator represents the total number of the positive sample and the negative sample. Wherein, the coefficient of the numerator is 2, because repeated calculation and common elements are involved in the denominator.

When selecting the loss function of the auxiliary task, CrossEntropy [48] Loss function is used as the auxiliary loss function. Its definition is as follows:

$$\text{CEL} = \sum_{i=1}^{N} y(i) \log \hat{y}(i) + (1 - y(i)) \log (1 - \hat{y}(i))$$ \hspace{1cm} (18)

where $i$ represents the $i$ sample, $N$ represents the Batch Size, $y(i)$ represents the semantic constraint angle of the predicted value of the sample, and $\hat{y}(i)$ represents the semantic constraint angle of the label value of the sample. The positive is 1, and the negative is 0.

IV. EXPERIMENTAL VERIFICATION

A. BRIEF DESCRIPTION OF DATASETS

To verify the effectiveness of the method proposed herein, two different datasets are designed to train and verify our model. One dataset is an open Massachusetts dataset, and the other is the private dataset marked and extracted from Google Map and is named RSR (Remote Sensing Rod) dataset. Besides, the deep learning framework used is [49]. Two datasets are briefly described as follows:

(1) Massachusetts road dataset. This dataset is composed of the aerial imagery of Massachusetts, including all areas in Massachusetts, such as cities, rural areas and suburbs. It covers an area of more than 2,600 km². The size of each original high spatial resolution remote sensing image is 1500 × 1500. There are 1,171 images. The resolution is 1m. The dataset is segmented into 1,108 training images, 49 test images and 14 verification images. The label image of this dataset is customized into a binary image. Wherein, roads are marked as the foreground, and other objects are marked as...
the background. As shown in Figure 9, the left and the right represent the original image and the label respectively.

(2) RSR dataset: Since roads in the Massachusetts dataset are not diversified and abundant, some sites are extracted from open map data on Google Map, including Tokyo in Japan, Washington in America, Xinjiang in China, London in Britain, etc. Kinds of road scenarios are designed, including urban roads, rural roads, desert roads, etc, as shown in Figure 10. The pixel size of each remote sensing image is 1024 × 1024. There are 450 training images and 50 test images.

B. EVALUATION METHOD

Precision, Recall, F1 Score and MIOU are used as the evaluation indexes of road extraction in this article. Their definitions are as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (19) \\
\text{Recall} = \frac{TP}{TP + FN} \quad (20) \\
\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (21)
\]

where TP is True Positive: the prediction is positive, and the ground truth result is positive too; FP is False Positive: the prediction is positive, and the ground truth result is negative; FN is False Negative: the prediction is negative, and the ground truth result is positive; TN is True Negative: the prediction is negative, and the ground truth result is negative too. As the result of F1 is greater, it means that the prediction image is more similar to the ground truth provided.

IOU (Intersection over Union) is used to evaluate the precision of pixel marked in image segmentation and has been used as a standard measure in semantic segmentation. The calculation is as follows:

\[
\text{IOU} = \frac{1}{2} \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{p_{ij}}{p_{ii} + \sum_{j=0}^{1} p_{ji} - p_{ii}} \quad (22)
\]

where \(i, j = 0\) represents the background and \(i, j = 1\) represents the road, \(p_{ij}\) represents the number of pixels which should belong to \(i\) type but are predicted as \(j\) type, \(p_{ji}\) represents the number of pixels which should belong to \(j\) type but are predicted as \(i\), and \(p_{ii}\) represents the number of correct predictions.

C. ANALYSIS OF THE RESULTS OF THE COMPARISON EXPERIMENT

In order to improve the generalization ability of the network, we use horizontal flipping, image random rotation, image mirroring flipping, HSV space transformation for data enhancement. HSV transformation is a means of data set amplification, where H represents hue, S represents saturation, and V represents brightness. HSV generally can eliminate the impact of lighting, brightness, and color differences on the picture. And the purpose of HSV transformation is also to pass the existing data to obtain more training data, and reduce the under-fitting or over-fitting problems caused by the data quality or the amount of data is too small. A training strategy with 300 epochs and a batchsize of 8 is used. a stochastic gradient descent (SGD) optimizer with a momentum of 0.9, a weight attenuation of 0.0005, and an initial learning rate of 0.003 are set in the experiment. And after every 100 epochs, the learning rate will automatically drop 10 times.

To compare the superiority of our method, UNet, LinkNet, DLinknet, BDNet and GLNet are used to compare with our model. And some methods with complex and deep models, like Exfuse, BiseNet, Context Encoding Network (EncNet), learning a discriminative feature network (DFN) and/or multi-scale context intertwining (MSCI), are also used to compare with our proposed method. The experimental results on the private dataset and the public dataset are as shown in Table 1 and Table 2.

When using the loss function mentioned in this paper, at the end of training, the CE and DL values of the loss function of each model are as follows: UNet: 0.069, 0.204; LinkNet:0.066, 0.201; DLinknet:0.059, 0.176; BDNet: 0.056, 0.174; GlNet: 0.061,0.177; Ours: 0.051, 0.158. The road extraction accuracy of the proposed model through the multi-task key point constraint module, the dual attention mechanism module, and the multi-branch cascading hollow space pyramid module, so that the CE and DL indicators in this paper are the lowest of all the comparison models.

Results in Table 1 and Table 2 reveal that our method is superior to compared models and realizes the optimal performance. In private dataset and public dataset, results of key indexes are as follows: IoU: 69.32% and 60.32%; F1: 81.43% and 75.17%; precision: 77.14% and 74.30%; recall: 86.23% and 76.10%.

For the comparison model shown in Table 1 and Table 2, UNet and LinkNet were early segmentation models that used encoding & decoding networks to achieve target segmentation. But the model structures of UNet and LinkNet...
TABLE 1. Comparison of experimental results of other road extraction models on RSR dataset.

| Model   | IoU   | F1    | Precision | Recall  | Fps/ms |
|---------|-------|-------|-----------|---------|--------|
| UNet    | 64.52 | 77.96 | 71.30     | 83.25   | 17.6   |
| LinkNet | 65.68 | 78.92 | 74.08     | 84.44   | 22.5   |
| DLNet   | 67.63 | 80.34 | 75.28     | 86.13   | 14.7   |
| BDNet   | 67.91 | 80.57 | 75.89     | 85.87   | 12.3   |
| GLNet   | 67.15 | 79.81 | 75.76     | 84.32   | 9.8    |
| Exfuse  | 67.86 | 80.49 | 75.71     | 85.91   | 6.9    |
| BiseNet | 66.19 | 79.38 | 74.93     | 84.39   | 9.2    |
| EncNet  | 65.82 | 79.11 | 75.63     | 82.93   | 13.3   |
| DFN     | 66.77 | 79.70 | 75.58     | 84.30   | 9.6    |
| MSCI    | 65.97 | 79.03 | 75.21     | 83.26   | 10.4   |
| Proposed| 69.32 | 81.43 | 77.14     | 86.23   | 7.3    |

TABLE 2. Comparison of experimental results of other road extraction models on massachusetts road dataset.

| Model   | IoU   | F1    | Precision | Recall  | Fps/ms |
|---------|-------|-------|-----------|---------|--------|
| UNet    | 57.22 | 72.38 | 72.15     | 72.61   | 13.6   |
| LinkNet | 57.79 | 72.87 | 71.37     | 74.43   | 18.9   |
| DLNet   | 59.06 | 74.20 | 72.98     | 75.46   | 11.2   |
| BDNet   | 59.46 | 74.26 | 73.55     | 74.98   | 7.1    |
| GLNet   | 59.33 | 74.21 | 72.60     | 75.89   | 6.7    |
| Exfuse  | 58.73 | 73.93 | 73.10     | 74.78   | 3.6    |
| BiseNet | 59.26 | 74.14 | 72.85     | 75.48   | 6.8    |
| EncNet  | 58.30 | 73.64 | 73.68     | 73.60   | 9.1    |
| DFN     | 58.65 | 73.77 | 72.89     | 74.67   | 5.2    |
| MSCI    | 57.82 | 72.91 | 72.12     | 73.72   | 7.7    |
| Proposed| 60.32 | 75.17 | 74.30     | 76.10   | 4.4    |

are simple and their results are worse than the results of other compared models both on RSR and Massachusetts Road Datasets. DLNet improves the accuracy of road extraction by adding a dilated convolution structure to LinkNet to increase the receptive field of the feature map, and the IoU, F1 score, Precision, and Recall are 67.63%, 80.34%, 75.28% and 86.13% on RSR Dataset. BDNet also adopts this dilated convolution structure. So the results of BDNet and DLNet are very close on both RSR and Massachusetts Road Datasets. GLNet is a dense network model based on Densenet and improves the road extraction effect by increasing the complexity of the network and achieves good effects, the IoU, F1 score, Precision and Recall are 67.15%, 79.81%, 75.76% and 84.32% on RSR Dataset, but the GLNet also involves a large quantity of calculation and the Fps is 9.8ms.

We also do some comparative experiments with the latest complex models, such as Exfuse, BiseNet, EncNet, DFN and MSCI. The experimental results are shown in Table 1 and Table 2. It can be seen that our model has better performance in terms of Precision, Recall, F1 Score and IOU. The complex deep learning models are proposed for conventional segmentation tasks, such as PASCAL VOC 2012 dataset, COCO dataset, PASCAL-Context and NYUv2, etc. The characteristics of the segmentation object and the background are different. For the road extraction in remote sensing image, the difficult point of road segmentation is that the road structure is complicated.

Based on Linknet, the proposed model not only adopt dilated convolution, but also combine branches of dilated convolution and ordinary convolution, integrate information of high-dimensional feature and low-dimensional feature and adds road constraint angle and PCA to further improve the effect. Therefore, our method has great strengths in terms of IoU, F1, Pr, Re, etc.

In addition, we take Fps as an indicator of computational cost. Because Linknet is the basic network for road segmentation, the Fps of Linknet can be Up to 22.5ms on the RSR data set. Compared with Linknet, Unet uses dimensional superposition in feature fusion, which increases the testing time. Based on Linknet, DLNet uses hole convolution to improve the road segmentation accuracy, but the increase in calculation parameters will also add calculation time. Compared with DLNet, although BDNet reduces the calculation parameters by constructing D-Blocks, it also increases the network depth and the fps of BDNet will also increase. GlNet uses a dense connection method, and the network depth is the deepest among these models, so the computational cost is relatively high. However, the Fps of the complex deep learning models is equivalent to our proposed model. The fps of Exfuse, BiseNet, EncNet, DFN, and MSCI is 3.6 ms, 6.8 ms, 9.1 ms, 5.2 ms, 7.7 ms, respectively. And the fps of our model is 4.4 ms.

Figure 11 and figure 12 displays the segmentation result of different roads. Wherein, the second row and the fourth row are the local amplification results of the first row and the third row. It can be seen that the method proposed in this article can solve the problem of false and missing detection of roads with a small pixel width proportion, improve performance of segmentation during road extraction and improve comprehensive evaluation indexes.

From the second row of Figure 11(c)-(f), it can be seen that BDNet, DLinknet, GLNet and LinkNet are more likely to wrongly recognize something non-road like shadows or junctions of different objects as roads. And in figure 11(g), the segmentation results of UNet even have omissions, resulting in road disconnection. Only the road extracted by our model is closest to the label image. For the results in the fourth row in Figure 11, the advantages of our method are much obvious. It can be seen that other models have generated some non-existent roads.

In Figure 12, it can be seen that the segmentation results of the models Exfuse, DFN, EncNet, and MSCI produce road recognition errors. Although BiseNet has no wrong roads, the segmented roads are blurry at the connection and even broken. Compared with other methods, it can be seen that only the road extracted by our model is clear and continuous.

In conclusion, the good performance of the method proposed in this paper on these two datasets reveals that our model is of good robustness in terms of road extraction from high spatial resolution remote sensing images and proves the effectiveness and reliability of the model in road extraction.

To further verify the effectiveness of the method proposed in this article, in verification with remote sensing road images.
in different scenarios, test images of the RSR dataset are classified into the non-shading scenario (N), general shading scenario (S) and multi-shading scenario (M) according to the complexity of scenarios in images. In the N scenario, roads are highlighted. In the general shading scenario, some positions of roads are shaded by buildings, trees and shadows, which will affect the road extraction effect. In multi-shading scenarios, a very long part of roads is covered by other objects, such as trees or shadows, so roads are very hard to be extracted.

Next, we only compare the results of the UNet, LinkNet, DLinknet, BDNet and GLNet. Partial experimental comparison results of three scenarios are as shown in Table 3 and Figure 14. Wherein, N_IoU represents IoU index in N scenario, F1_N represents F1 index in N scenario and so on. Mean_IoU and Mean_F1 represent the mean value of F1 of
TABLE 3. Experimental results of different models in different scenarios in the private dataset.

| Model      | N IoU | S IoU | F1 N | S F1 | IoU  | F1 M | Mean IoU | Mean F1 |
|------------|-------|-------|------|------|------|------|----------|---------|
| Unet       | 72.32 | 64.06 | 77.85| 59.02| 73.90| 64.52| 77.96    |
| LinkNet    | 74.29 | 66.31 | 79.62| 58.74| 73.81| 65.68| 78.92    |
| DLinkNet   | 74.77 | 68.51 | 81.14| 61.62| 75.99| 67.63| 80.34    |
| BDNet      | 74.20 | 68.40 | 81.17| 62.83| 76.82| 67.91| 80.57    |
| GLNet      | 74.28 | 66.96 | 79.73| 63.23| 75.88| 67.15| 79.81    |
| Proposed   | 74.83 | 69.51 | 81.95| 64.88| 78.36| 69.32| 81.43    |

FIGURE 13. Results of Comparison between the Method Proposed and Existing Methods. (a)Input image; (b)Label image; (c)BDnet; (d)Dlinknet; (e)GLnet; (f)Unet; (g)Proposed.

FIGURE 14. The results for PCA on the Private Dataset (a) Original image (b) Label (c) Linknet (d)Linknet+PA (e) Linknet+CA (f)Linknet+PCA.

three scenarios. Results show that the IoU, F1, precision and recall of our model in three scenarios are superior to those of other compared models, and the mean IoU and mean F1 score is 69.32% and 81.43% respectively. Even compared with the compared model which has the best performance, the mean IoU and mean F1 of our model in three scenarios increase by 1.4% and 0.9% respectively.

The road extraction results are as shown in Figure 13. The first row is road extraction in N scenario, the second row is road extraction in S scenario, and the third row is the experimental result in M scenario. From the results of the first and second rows, it can be seen that for some unobvious roads, although some other models can also extract roads more or less, the extraction effect is not as good as the model we proposed. In terms of clarity, road continuity and ductility are also the best results of our model. Especially, from the results of the third row, it can be seen that none of the other models can detect the section of road covered by trees, resulting in obvious disconnection in the extraction results. our proposed model can detect this section of the road, which proves the superiority of our model.

D. ANALYSIS OF RESULTS OF ABLATION EXPERIMENT

To verify the superiority of our multi-task IPC, PCA and multi-branch CDSP in terms of road extraction and segmentation, an ablation experiment is conducted for three modules in this article.

1) MULTI-TASK IPC

The superiority of the IPC module in road extraction was verified on the private dataset and the public dataset respectively. Experimental results are shown in Table 4 and Table 5. In the private dataset, IoU and F1 of LinkNet with IPC module increase by 3.0% and 2.2% respectively than that without IPC module. In the public dataset, IoU and F1 of LinkNet with IPC module increase by 1.8% and
TABLE 4. Comparison of results of multi-task IPC on RSR dataset.

| Model       | IoU  | F1   | Precision | Recall |
|-------------|------|------|-----------|--------|
| LinkNet     | 65.68| 78.92| 74.08     | 84.44  |
| LinkNet+IPC | 68.65| 80.41| 75.65     | 85.81  |

TABLE 5. Comparison of results of multi-task IPC on Massachusetts Road dataset.

| Model       | IoU  | F1   | Precision | Recall |
|-------------|------|------|-----------|--------|
| LinkNet     | 57.79| 72.87| 71.37     | 74.43  |
| LinkNet+IPC | 59.58| 73.93| 72.62     | 75.28  |

1.8% respectively than that without IPC module. Results show that multi-task IPC can improve the performance of road extraction models significantly, because IPC module can make use of road intersection, the prior knowledge, effectively.

2) PCA
To verify the impacts of PCA on road extraction of models, an ablation experiment was carried out for the PA mechanism and CA mechanism respectively. Results on the private dataset and the public dataset are as shown in Table 6 and Table 7. Results show that the performance in both the private dataset and the public dataset is improved by adding PA or CA, but the experimental effect is the best when both the PA and the CA are added. Compared with ordinary Linknet, when both two modules are added, IoU increases by 1.8%, and F1 increases by 1.3% on the private dataset. IoU increases by 1.6% and F1 increases by 1.4% on the public dataset. It verifies the effectiveness of PCA in road extraction and further indicates that the fusion of semantic information of images can improve the precision of road extraction.

The visualization results for PCA on the RSR Dataset and Massachusetts Road Dataset are shown in Figure 14 and Figure 15.

From the extracted road results, it can be seen that in the segmentation results of the LinkNet model, the roads generated in the area marked by the red circle are very blurred and generate a lot of noise, as shown in Figure 14(c). After adding the position attention mechanism and the channel attention mechanism respectively, the segmented roads with Linknet+PA and Linknet+CA are smoother, as shown in Figure 14(d) and (e), but there will still be wrongly segmented roads or disconnected roads. Only when the position attention mechanism and the channel attention mechanism are added at the same time, the result of Linknet+PCA does not generate the wrong road and the extracted road is closest to the label image. The superiority of the PCA module is proved by the road extraction results on the private data set. Similar results can also be verified on the public data set, as shown in Figure 15.

3) MULTI-BRANCH CDSP
To verify the role of multi-branch CDSP in performance improvement of models, on the basis of multi-task IPC and PCA, results before and after stacking of multi-branch CDSP are compared. The comparison results are shown in Table 8 and Table 9.

Results show that compared with models without adding multi-branch CDSP, indexes on both the public dataset and the private dataset are improved. The improvement effect of this module is limited, but multi-branch CDSP can enhance the ability to restore information via feature fusion.
Comprehensive results indicate that compared with the basic model LinkNet (See the results in Table 6–7), IoU and F1 on the private dataset increase by 3.6% and 2.5%, respectively. IoU and F1 on the public dataset increase by 2.5% and 2.3% respectively, further proving the effectiveness of the method proposed in this article.

V. CONCLUSION

To improve LinkNet, a network model for extraction of target roads from multi-scenario remote sensing images is proposed in this article. To solve the problem of unstable connection between road intersections in high spatial resolution remote sensing images, failure of effective fusion of semantic information and loss of information during down-sampling, multi-task IPC, PCA and multi-branch CDSP are designed based on LinkNet respectively. Model analysis and experimental results indicate that all the models designed in this article can improve the final performance of road extraction, especially in many scenarios, such as coverage, shading, etc. The method improves the performance of road extraction from remote sensing images to some extent, but the following aspects are to be improved: (1) due to the auxiliary task of multi-task key point constraints, the network needs to calculate the skeleton information of the road label map additionally; (2) the subsequent burning algorithm and pathfinding algorithm will increase the computational cost; (3) the multiple scales of void convolution in the multi-branch cascaded void space pyramid module will also inevitably increase the calculation parameters of the network and increase in time cost. (4) It has been proven that our models improve the precision of road extraction, since the angle and other parameters are added during road extraction, multi-branch and multi-scale road feature are fused, road features of channels and positions are extracted, the corresponding road extraction time of networks will be prolonged. We need to explore some new methods which can minimize the road extraction time without affecting the accuracy of models.

REFERENCES

[1] S. Abdullahi, B. Pradhan, and M. N. Jebar, “GIS-based sustainable city compactness assessment using integration of MCDM, Bayes theorem and RADAR technology,” Geocarto Int., vol. 30, no. 4, pp. 1–39, 2015.
[2] A. M. Youssef, S. A. Sefry, B. Pradhan, and E. A. Alfaadil, “Analysis on causes of flash flood in Jeddah City (Kingdom of Saudi Arabia) of 2009 and 2011 using multi-sensor remote sensing data and GIS,” Geomatics, Natural Hazards Risk, vol. 7, no. 3, pp. 1018–1042, 2016.
[3] Q. Weng, “Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends,” Remote Sens. Environ., vol. 117, pp. 34–49, Feb. 2012.
[4] J. Cheng, W. Ding, X. Ku, and J. Sun, “Road extraction from high-resolution SAR images via automatic local detecting and human-guided global tracking,” Int. J. Antennas Propag., vol. 2012, no. 10, pp. 1497–1500, 2014.
[5] Y. Li, L. Xu, J. Rao, L. Gao, Z. Yan, and S. Jin, “A Y-Net deep learning method for road segmentation using high-resolution visible remote sensing images,” Remote Sens. Lett., vol. 10, no. 4, pp. 381–390, Apr. 2019.
[6] P. Shamsoolmoli, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, “Road segmentation for remote sensing images using adversarial spatial pyramid networks,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 6, pp. 4673–4688, Jun. 2021.
[7] B. Parajuli, P. Kumar, T. Mukherjee, E. Passiliao, and S. Jambawalikar, “Fusion of aerial lidar and images for road segmentation with deep CNN,” in Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., Nov. 2018, pp. 548–551, doi: 10.1145/3274895.3274993.
[8] M. Amo, F. Martinez, and M. Torre, “Road extraction from aerial images using a region competition algorithm,” IEEE Trans. Image Process., vol. 15, no. 5, pp. 1192–1201, May 2006.
[9] M. Song and D. Civco, “Road extraction using SVM and image segmentation,” Photogramm. Eng. Remote Sens., vol. 70, no. 12, pp. 1365–1371, Dec. 2004.
[10] P. H. Chen, C. J. Lin, and B. Schikopf, “A tutorial on V-support vector machines,” Appl. Stochastic Models Bus. Ind., vol. 21, no. 2, pp. 111–136, 2005.
[11] V. Mnih, Machine Learning for Aerial Image Labeling. Toronto, ON, Canada: Toronto Univ. Canada, 2013.
[12] V. Caselles, R. Kimmel, and G. Sapiro, “Geodesic active contour,” Int. J. Comput. Vis., vol. 22, no. 1, pp. 61–79, 1997.
[13] D. Chaudhuri, N. K. Kushwaha, and A. Samal, “Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques,” IEEE J. Sel. Topics Appl. Earth Observe. Remote Sens., vol. 5, no. 5, pp. 1538–1544, Oct. 2012.
[14] Z. Miao, B. Wang, W. Shi, and H. Zhang, “A semi-automatic method for road centerline extraction from VHR images,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 11, pp. 1856–1860, Nov. 2014.
[15] C. Ussalan and B. Sirmacek, “Road network detection using probabilistic and graph theoretical methods,” IEEE Trans. Geosci. Remote Sens., vol. 50, no. 11, pp. 4441–4453, Nov. 2012.
[16] T. Sun, Z. Chen, W. Yang, and Y. Wang, “Stacked U-Nets with multi-output for road extraction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 202–206.
[17] V. Mnih and G. E. Hinton, “Learning to detect roads in high-resolution aerial images,” in Proc. Eur. Conf. Comput. Vis. Workshops, 2015, pp. 371–379,Mar. 2007.
[18] S. Saito, T. Yamashita, and Y. Aoki, “Multiple object extraction from aerial imagery with convolutional neural networks,” Electron. Imag., vol. 2016, no. 10, pp. 1–9, Feb. 2016.
[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 30–42, Jan. 2012.
[20] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[21] M. Ranzato, Y. Boureau, S. Chopra, and Y. LeCun, “A unified energy-based framework for unsupervised learning,” J. Mach. Learn. Res., vol. 2, pp. 371–379, Mar. 2007.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” 2014. arXiv:1409.0575. [Online]. Available: http://arxiv.org/abs/1409.0575
[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[25] 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/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 182–186.
[26] A. Chaurasia and E. Culurciello, “LinkNet: Exploiting encoder representation for efficient semantic segmentation,” in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2017, pp. 1–4.
[27] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in Proc. Conf. Neural Inf. Process. Syst., Barcelona, Spain, vol. 2016, pp. 4988–4996.
[28] Y. Xu, Z. Peng, L. He, K. Fan, Z. Li, and L. Tong, “Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks,” Sensors, vol. 19, no. 19, p. 4115, Sep. 2019, doi: 10.3390/s19194115.
[29] R. Kestur, S. Farooq, R. Abdal, E. Mehraj, O. S. Narasipura, and M. Mudigere, “UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle,” J. Appl. Remote Sens., vol. 12, no. 1, pp. 1–10, 2018.
[30] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[31] V. Vaquero, I. del Pino, F. Moreno-Noguer, J. Sola, A. Sanfeliu, and J. Andrade-Cetto, “Decoronvolutional networks for point-cloud vehicle detection and tracking in driving scenarios,” in Proc. Eur. Conf. Mobile Robots (ECMR), Sep. 2017, pp. 1–7.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[33] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “BiSeNet: Bilateral segmentation network for real-time semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Munich, Germany, 2018, pp. 334–349.

[34] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 1857–1866.

[35] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7151–7160.

[36] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, “Enhancing feature fusion for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 269–284, doi: 10.1007/978-3-030-01249-6_17.

[37] L. Di, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, “Multi-scale context intertwining for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 622–638.

[38] Y. Cai, X. Wang, Z. Yu, F. Li, P. Xu, Y. Li, and L. Li, “Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network,” IEEE Access, vol. 7, pp. 183706–183716, 2019, doi: 10.1109/ACCESS.2019.2958864.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2261–2269.

[41] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. Berlin, Germany: Springer, 2017, doi: 10.1007/978-3-662-54345-0_3.

[42] R. Cui and M. Liu, “Hippocampus analysis by combination of 3-D DenseNet and shapes for Alzheimer’s disease diagnosis,” IEEE J. Biomed. Health Inform., vol. 23, no. 5, pp. 2099–2107, Sep. 2019.

[43] T. Y. Zhang and C. Y. Suen, “A fast parallel algorithm for thinning digital patterns,” Commun. ACM, vol. 27, no. 3, pp. 236–239, Mar. 1984.

[44] S. M. Persson and I. Sharf, “Sampling-based A* algorithm for robot path planning,” Int. J. Robot. Res., vol. 33, no. 13, pp. 1683–1708, Nov. 2014.

[45] H. J. Mattausch, N. Omori, S. Fukae, T. Koide, and T. Gyoten, “Fully-pattern-matching engine with dynamic adaptability to Hamming or manhattan distance,” in Symp. VLSI Circuits. Dig. Tech. Papers, Jun. 2002, pp. 252–255.

[46] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Proc. 32nd Int. Conf. Int. Conf. Mach. Learn., Jun. 2015, pp. 448–456.

[47] T. Chen and X. Wu, “Characteristics of activation function in sigma-pi neural networks,” J. Fudan Univ., vol. 36, no. 6, pp. 639–644, 1997.

[48] D. E. Booth, “The cross-entropy method,” Technometrics, vol. 50, no. 1, p. 92, 2008.

[49] J. Rauber, R. Zimmermann, M. Bethge, and W. Brendel, “Footboolx native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX,” J. Open Source Softw., vol. 5, no. 53, p. 2607, Sep. 2020.

[50] XUNGEN LI received the B.E. degree from the China University of Petroleum, Qingdao, in 1995, and the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2005.

He is currently an Associate Professor with Hangzhou Dianzi University. He is also with the Key Laboratory of Radio Frequency Circuit and System of Ministry of Education, Hangzhou. His research interests include integrated circuit and artificial intelligence, deep learning, and machine learning.

[51] ZHAN ZHANG received the B.E. degree from Shandong Jiaotong University, in 2016. He is currently pursuing the degree with the College of Electronics and Information, Hangzhou Dianzi University. His research interests include image processing and deep learning.

[52] SHUAIHUAI LV received the B.E. and Ph.D. degrees from the School of Automation, Northwestern Polytechnical University, in 2011 and 2016, respectively. Since 2016, he has been a Lecturer with the College of Electronics and Information, Hangzhou Dianzi University. His research interests include deep learning, target detection, network control systems, and consensus of distributed multi-agent system and its applications.

[53] MIAN PAN received the B.E. and Ph.D. degrees from the School of Electronic Engineering, Xidian University, in 2007 and 2013, respectively. Since 2013, he has been a Lecturer with the College of Electronics and Information, Hangzhou Dianzi University. His research interests include adaptive signal processing, statistical signal processing, and machine learning and its applications.

[54] QI MA received the B.E. and Ph.D. degrees from the Department of Information Science and Electronic Engineering, Zhejiang University, in 1990 and 2000, respectively. He is currently a Professor with the College of Electronics and Information, Hangzhou Dianzi University. His research interests include VLSI/SoC design, IC design methodology, intelligent processing of audio/voice and machine learning algorithm of arc fault detection, and load identification.

[55] HABIN YU received the Ph.D. degree in information and communication engineering from Zhejiang University, in 2007. From November 2016 to November 2017, he was a Visiting Scholar with the University of Pittsburgh. He is currently a Professor with the College of Electronics and Information, Hangzhou Dianzi University. His research interests include computer vision, image processing, machine learning, multi-sensor fusion, and marine electronic information systems.