DCTNET: HYBRID NETWORK MODEL FUSING WITH MULTISCALE DEFORMABLE CNN AND TRANSFORMER STRUCTURE FOR ROAD EXTRACTION FROM GAOFEN SATELLITE REMOTE SENSING IMAGE

Qinglie Yuan

School of Civil and Architecture Engineering, Panzhihua University, Panzhihua 617000, China- yuanqinglie@pzhu.edu.cn.

KEY WORDS: road extraction, Transformer, convolutional neural network, deep learning

ABSTRACT:

The urban road network detection and extraction have significant applications in many domains, such as intelligent transportation and navigation, urban planning, and automatic driving. Although manual annotation methods can provide accurate road network maps, their low efficiency with high-cost consumption are insufficient for the current tasks. Traditional methods based on spectral or geometric information rely on shallow features and often struggle with low semantic segmentation accuracy in complex remote sensing backgrounds. In recent years, deep convolutional neural networks (CNN) have provided robust feature representations to distinguish complex terrain objects. However, these CNNs ignore the fusion of global-local contexts and are often confused with other types of features, especially buildings. In addition, conventional convolution operations use a fixed template paradigm to aggregate local feature information. The road features present complex linear-shape geometric relationships, which brings some obstacles to feature construction. To address the above issues, we proposed a hybrid network structure that combines the advantages of CNN and transformer models. Specifically, a multiscale deformable convolution module has been developed to capture local road context information adaptively. The Transformer model is introduced into the encoder to enhance semantic information to build the global context. Meanwhile, the CNN features are fused with the transformer features. Finally, the model outputs a road extraction prediction map in high spatial resolution. Quantitative analysis and visual expression confirm that the proposed model can effectively and automatically extract road features from complex remote sensing backgrounds, outperforming state-of-the-art methods with IOU by 86.5% and OA by 97.4%.

1. INTRODUCTION

With the rapid development of urbanization, the construction of smart cities has risen extensive attention. Urban infrastructure's geographic information updating has greatly improved (Yuan et al., 2021). In city planning, extracting road elements has become an important component. Road features from high-resolution remote sensing images can provide rich and accurate spatial information to urban construction, making sufficient preparations for updating the city information database.

The study of road extraction can provide scientific decisions for urban planning, management, and decision-making. Roads are the main component of urban transportation, essential elements of urban geographic information, and cultural and economic exchange hubs. That makes the extraction of road information of great practical significance. In recent years, extensive research has been conducted on road extraction from high-resolution remote sensing images, and various methods have been proposed for different application fields of road information. However, there are many challenges in extracting road information, such as complex shapes, similar spectral features, building shadow, or tree occlusion, as shown in Figure 1.

Traditionally, the main methods using high-resolution remote sensing images include pixel-based and object-oriented methods.

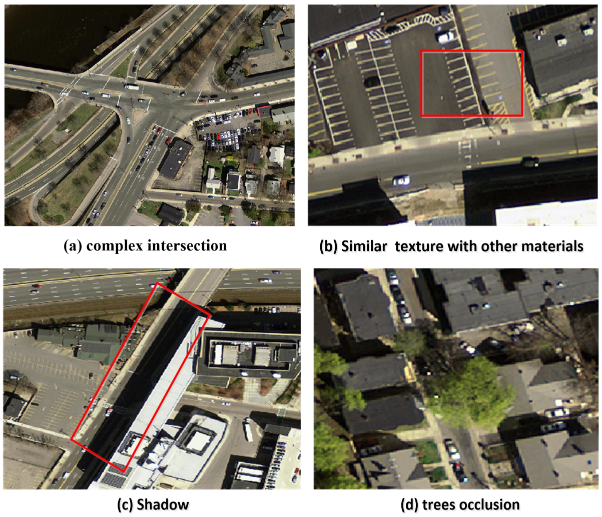


Figure 1. Some challenges in road extraction.

In the pixel-based methods, the spectral and texture features or geometric topology of the road itself are utilized to extract the road skeleton by template matching or knowledge-driven. The former has good road extraction performance, but the manual intervention is high, with much seed point selection and low processing automation. The latter has low human involvement, but this algorithm has high computational complexity and low operational efficiency (Zhu et al., 2021). The object-based method regards the road area as a whole and extracts the road

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-M-3-2023-273-2023

information in the high-resolution image by image segmentation clustering, support vector machine, conditional random field, and other algorithms (Yuan et al., 2021). However, these methods have complex operations and poor stability and are affected by uneven variation in the grayscale of image pixels. The extraction results have problems such as breakage, incorrect extraction, missing structure, etc. The road extraction effectiveness is poor, and the post-processing work is time-consuming and laborious, making it difficult to apply to large-scale remote sensing scenes. Therefore, there is an urgent need for a fast and accurate method for extracting roads from high-resolution remote sensing images automatically (Gao et al. 2018).

In recent years, deep learning has been widely applied in various tasks of remote sensing images and has achieved significant results in the field of computer vision (Yuan et al., 2022). Many road extraction models have been proposed based on deep learning. Mnih et al. (2010) used deep learning for road extraction tasks, and the results showed that deep learning methods could complete road extraction tasks and achieve better results than traditional methods. For example, Zhong et al. (2017) applied the fully convolutional neural network (FCN) to the road extraction task, achieving accurate results. Chen et al. (2016) proposed a SegNet-based network structure using encoders and decoders to utilize contextual information and preserve image contour details.

In road extraction from high-resolution remote sensing images, Mendes et al. (2016) proposed a network model that can extract complete and clear edges of roads from high-resolution images, but it is difficult to distinguish road types accurately. Cheng et al. (2017) proposed CasNet Network utilizing two cascaded CNNs to achieve road detection and extraction, but the extraction results in occluded areas are poor. Almeida et al. (2020) combined multiple deep neural networks to construct an ENet model for road detection. Chaurasia et al. (2017) used a lightweight ResNet18 encoder to build a semantic segmentation model, Link Net, which ensures road detection accuracy and improves road detection efficiency. Zhou et al. (2018) proposed an improved Dilation Convolution Link Net (DLinkNet) model. Based on the DLinkNet model, the receptive field is enlarged by using dilated convolution to obtain rich contextual semantic information, which improves the road integrity and edge clarity, but the extraction is poor for the small, shaded, and crossed roads.

Although CNN-based methods have achieved excellent performance in image segmentation, they still cannot be competent with the requirements for segmentation accuracy in road feature extraction. Road segmentation using remote sensing images remains a challenging task due to the inherent locality of convolution operations. It is difficult for CNN-based methods to learn global semantic information with relatively long distances in pairing pixels. Some methods attempt to solve this problem through dilated convolutions, self-attention mechanisms, and feature pyramids (Lin et al., 2017). However, these methods still have limitations in establishing long-distance dependencies.

CNN has translation invariance and local sensitivity. Convolutional kernels can capture the fine-grained features and local information of roads. However, CNN's receptive field is limited, and it cannot obtain global information. And with the improvement of computing power, the demand for data in CNN models is becoming increasingly saturated, requiring larger models to replace CNN. Thus, inspired by the great success of natural language processing (NLP), Transformer gradually began to be applied in Computer Vision (CV). For example, The Image Transformer model (Parmar et al., 2018) applies Transformer to computer vision. The proposal of the object detection model DETR (Carion et al., 2020) and the image classification model ViT (Dosovitskiy et al., 2021) has promoted the rapid development of visual transformers. However, due to many parameters and the high computational cost of the Transformer model, many methods began introducing prior knowledge from CNN into the Transformer, including locality, hierarchy, multiscale, residual connection, and bias design. Swin Transformer (Liu et al., 2021) has achieved advanced results in multiple visual tasks.

To address the above issues, we proposed a hybrid network structure DCTNet: a hybrid network model fusing with multiscale deformable CNN and Transformer structure for road extraction. The network model uses the GaoFen-6 satellite remote images as the training samples that contain various road scene types, such as urban traffic roads, internal roads of buildings, and rural roads.

2. THE PROPOSED NETWORK MODEL

2.1 The overall architecture of the model

As illustrated in Figure 2, DCTNet combines the advantages of CNN and transformer models. Specifically, a multiscale deformable convolution module has been developed to capture local road context information adaptively. In the encoders, this module constructs pooling layers with different scales and predicts the deformation parameters of convolution to extract local features of complex roads. The Transformer model is introduced into the encoder to enhance semantic information to build the global context. Meanwhile, the CNN features from the encoder are fused with the transformer features. Finally, the model outputs a road extraction prediction map in high spatial resolution.

2.2 CNN and Swin Transformer encoders

In this study, a dual branch encoder was established to extract road features. One branch uses a residual network structure. He Kaiming et al. (2016) proposed ResNet, which uses deep residual learning to solve training optimization problems such as gradient vanishing or explosion. Compared with ordinary networks, ResNet's residual blocks mainly introduced a skip connection between input and output, making it easier for the network to train and learn multi-level features. As shown in Figure 2, the structure of the ResNet residual block is shown, where X_i and X_{i+1} are the inputs of layer *i* and the outputs of

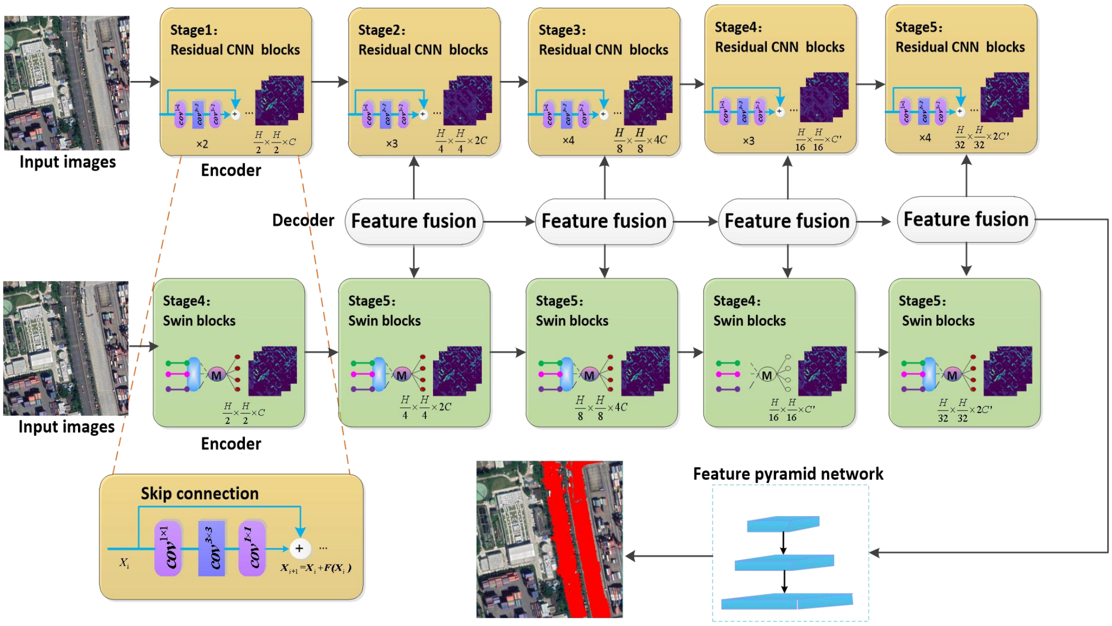

Input images
Stage1: Residual CNN blocks
Stage2: Residual CNN blocks
Stage3: Residual CNN blocks
Stage4: Residual CNN blocks
Stage5: Residual CNN blocks
×2
×3
×4
×3
×4
$\frac{H}{2}\times\frac{H}{2}\times C$
$\frac{H}{4}\times\frac{H}{4}\times 2C$
$\frac{H}{8}\times\frac{H}{8}\times 4C$
$\frac{H}{16}\times\frac{H}{16}\times C'$
$\frac{H}{32}\times\frac{H}{32}\times 2C'$
Encoder
Decoder
Feature fusion
Feature fusion
Feature fusion
Feature fusion
Input images
Stage4: Swin blocks
Stage5: Swin blocks
Stage5: Swin blocks
Stage4: Swin blocks
Stage5: Swin blocks
M
Encoder
Skip connection
X_i
$cov^{1\times1}$
$cov^{3\times3}$
$cov^{1\times1}$
$X_{i+1}=X_i+F(X_i)$
Feature pyramid network

Figure 2. The proposed network framework.

layer i+1, and $F(X_i)$ is the residual function. In this paper, ResNet34 is used as the backbone of the CNN branch, which consists of 5 residual blocks, where block1 consists of 7×7 convolution, input batch normalization layer (BN), ReLU layer, and 3×3 maximum pooling layers. The other blocks have a similar structure, mainly composed of multiple convolutional blocks, bottleneck structures, and residual connections. Finally, the feature maps can be obtained with original input sizes of 1/2, 1/4, 1/8, 1/16, and 1/32.

The other branch encoders adopted the Swin Transformer (Liu et al., 2021) structure. Compared to the ViT network structure, the Swin Transformer only operates on a computation within a 7×7 window, performing self-attention computation to reduce computational complexity. Moreover, this model uses padding methods to ensure that the window can evenly divide the image. Unlike traditional multi-head self-attention (MSA) modules, Swin Transform is constructed based on a shift window, which replaces traditional MSA with W-MSA, as presented in equations (1)~(4). As shown in Figure 3, Swin Transformer inputs feature maps into the encoder blocks, sequentially passing through LayerNorm (LN) layer, W-MSA, MLP layer, SW-MSA, and MLP layer. Compared to W-MSA, the advantage of SW-MSA lies in the execution of shift windows, which enable information exchange.

$$\hat{z}^{l} = W\text{-}MSA(LN(z^{l-1})) + z^{l-1} \tag{1}$$

$$z^{l} = MLP(LN(\hat{z}^{l}) + \hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1} = S\text{-}WMSA(LN(z^{l})) + z^{l} \tag{3}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1}) + \hat{z}^{l+1} \tag{4}$$

where, Z^l and z^{l+1} represent the outputs of the l layer; LN denotes layer normalization; MLP denotes the multi-layer perceptron.

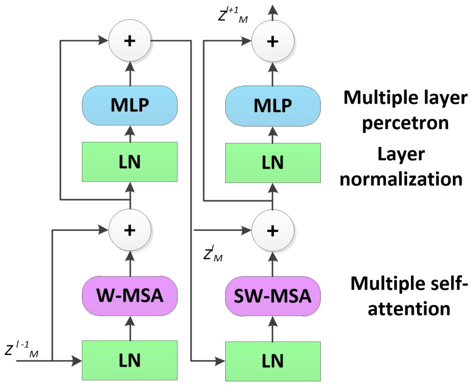


Figure 3. Swin Transformer block structure.

2.3 Multiscale Feature fusion decoders

The road presents various complex geometric shapes, but the convolution adopts a fixed-scale template calculation. To adaptively capture geometric structural features, the decoder introduces multiscale deformation convolution at the end of each residual block, as illustrated in Figure 4. Specifically, the feature maps are input into a multiscale pooling pyramid with a pooling kernel size of 3×3, 8×8, 64×64, 2×4, 4×2, 2×8, 8×2. Then, all features are upsampled using bilinear interpolation and concatenated via the channel. Meanwhile, deformation offsets are predicted via convolution 3×3. Finally, new features were constructed by deformation convolution and skip connections.

In feature fusion, CNN and Transformer features are input into the feature pyramid network. Specifically, Transformer features are reshaped into 2D images and fused with CNN features using addition operations and convolution 3×3. Then, from high to low levels, the feature maps are fused and upsampled by convolution 1×1 layer. Finally, classification prediction maps can be output with two channels to represent the probability of the background and road. The above process can integrate local features and global context to enhance semantic information and fine-grained spatial details.

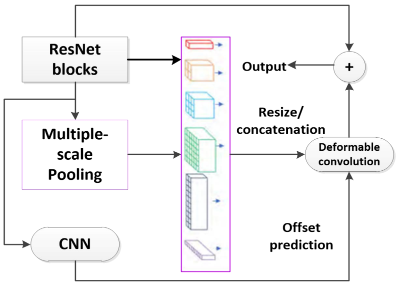


Figure 4. Multiscale deformable convolution head.

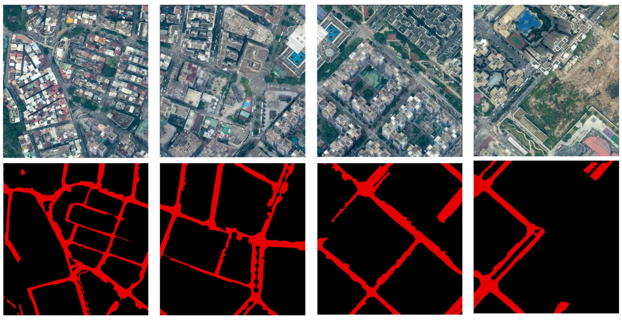
3. EXPERIMENTAL RESULTS AND DISCUSSION

This experimental configuration uses CPU i7-13700k and the GPU with GeForce RTX 4070Ti. The deep learning framework is Pytorch 1.7. This paper uses the Adam algorithm as the optimizer in the gradient descent process. Momentum is set to 0.9, and weight attenuation is set to 0.00005. The initial Learning rate of the experiment is set to 0.001. The piece-wise constant attenuation method is used. When epoch is 15, 30, and 50, the learning rate is set to 0.0008, 0.0006, and 0.0005 respectively. In the experiment, the batch size was set to 64, and the epoch was set to 120. In training, the image size was clipped to 512× 512. Data augmentation uses brightness transformation and Rotate 90°, mirror 180°. The dataset is divided into training and testing sets in an 8:2 ratio, and all experiments were loaded with pre-training weights on ImageNet.

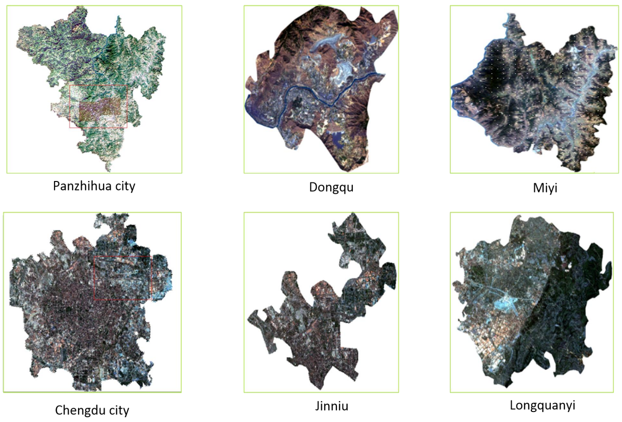
3.1 Data Description

Gaofen satellite road dataset covers urban and suburbs in Chengdu and Panzhihua, China, with a total coverage area of over 2000 km^2, as illustrated in Figure 5. The preprocessing of Gaofen images includes radiometric calibration, atmospheric correction, image fusion, and geometric correction. This dataset significantly differs in the image's road texture, color, environment, and imbalanced samples. Buildings and vegetation obstruct the road. The Gao-fen 6 satellite images include multispectral images (8m spatial resolution) with RGB and near-infrared bands and panchromatic images (2m spatial resolution). Finally, multispectral images with a spatial resolution of 2 meters were used in the dataset.

The images are cropped to 512×512 pixels without overlap. Datasets are randomly divided and selected as training samples 4200 and testing samples 600 and 1200 validation sets, respectively. Figure 5(a) presents some training samples and ground truth. Figure 5(b) presents the Gaofen images after preprocessing. The details in different areas are enlarged in the red rectangle.



(a)



(b)

Figure 5. The overview of the dataset.

3.2 Comparison with different methods

Some representative samples are visualized, as shown in Figure 6. Although the proposed method misclassifies some features in the test images, it outperforms other advanced models. For example, for suburban roads, as shown in the first row of the image, UNet and Deeplabv3+ (Chen et al., 2017) cannot segment complete road elements. In urban areas, these two methods cannot accurately distinguish similar textures. Some buildings were misclassified as roads. Segformer is very sensitive to buildings and ground and has many misclassifications. Moreover, it has poor discriminative ability for road details and geometric shapes. In contrast, Swin-Unet can extract complete road data, such as the results in the second row. However, some objects with similar textures still have weak predictive performance.

As reported in Table 1, quantitative results confirm that DCTNet can effectively and automatically extract road features from complex remote sensing backgrounds. DCTNet outperforms DeeplabV3+ methods with IOU by 1.34% and OA by 0.15% in the test1.

Test Datasets	**Test1**		**Test 2**	
Metrics	OA(%)	IOU(%)	IOU(%)	OA(%)
Swin-Unet	92.15	86.38	84.89	92.22
UNet	93.67	88.63	85.17	94.19
Segmenter	93.45	90.17	84.34	93.47
DeeplabV3+	94.52	91.06	84.28	96.82
DCTNet	**94.67**	**92.4**	**86.51**	**97.43**

Table 1. Comparison Accuracy using different methods. The bold values denote the best result.

As illustrated in the urban regions, the background environment of the road changes complexly, and the shadows of buildings and vegetation significantly obstruct the road, bringing challenges for road extraction. Test 2 of Table 1 shows that the Swin-Unet (Cao et al., 2022) extraction result is poor with 84.89% IOU, mainly since SwinUnet uses shallow features but ignores refinement of information, resulting in loss of details. In addition, due to the insufficient spatial consistency, poor continuity of road extraction results exists with fractures and omissions. The UNet (Siddique et al., 2021) model utilizes high-resolution features to improve boundary segmentation, but

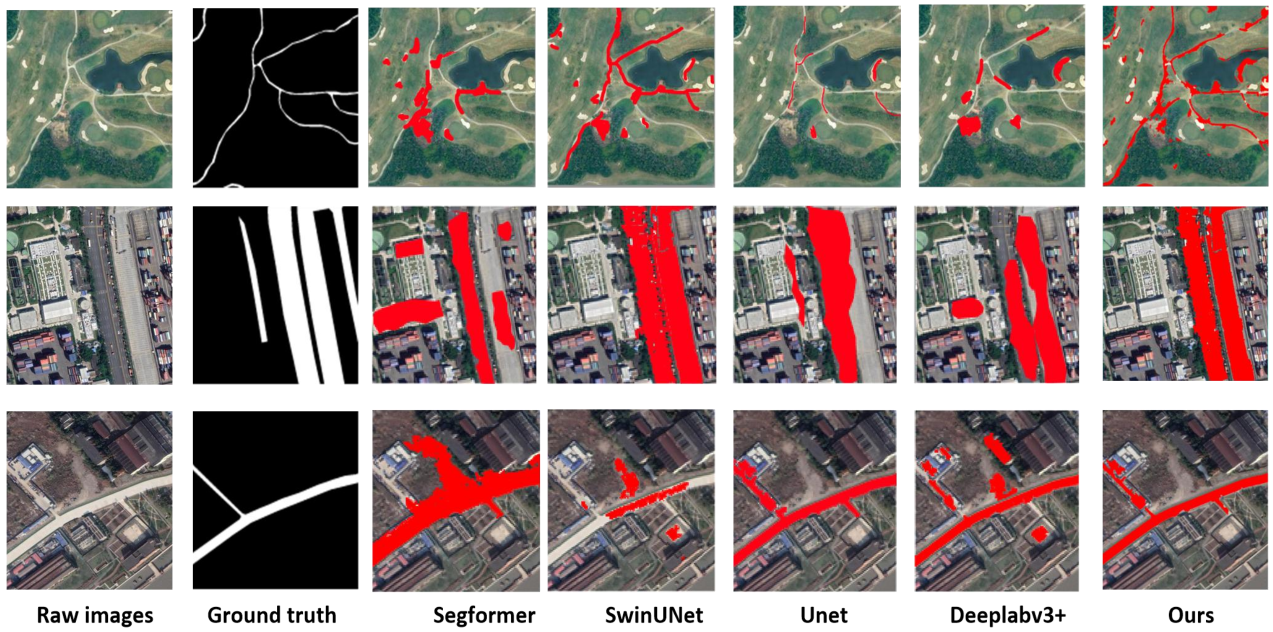


Figure 6. The road extraction results using different methods

the extraction of road details is insufficient, which triggers problems for small-scale road extraction. The DeepLabV3+ model introduces an ASPP module, which achieves multiscale feature fusion, improves boundary accuracy, and has good global integrity. However, the extraction is poor under shadows and occlusion areas.

The performance between Segmenter and DeepLabV3+ models in quantitative metric is not significant, but from the visualization results, the UNet model has obvious spectral texture and performs well in single road extraction. However, due to the model's simplicity, its performance in road extraction in complex environments is poor. The proposed model encoder adopts the hybrid network and introduces a cascaded feature fusion between the encoder and decoder, which outperforms the DeepLabV3+ network in road extraction under shadow and occlusion conditions. The visualization results of the model extraction in this article indicate that integrating local and global contexts can improve road accuracy.

4. CONCLUSION

This paper proposed a hybrid network model, DCTNet, that combines the advantages of CNN and Transformer to extract road information. To address the issues of insufficient contextual semantics and low extraction accuracy, the DCTNet improves the road extraction task using high spatial resolution optical remote sensing images. This network constructs dual branch encodes, which use the residual network and Swin-Transformer to generate local and global context dependencies for the overall road segmentation. The multiscale deformation convolution module can enhance the model's adaptive segmentation ability for complex shapes. The proposed model can effectively and automatically extract road features from complex remote sensing backgrounds, outperforming state-of-the-art methods with IOU by 86.5% and OA by 97.4%.

REFERENCES

Almeida T, Lourenco B, Santos V., 2020. Road detection based on simultaneous deep learning approaches. *Robotics and Autonomous Systems*, 133, 103605.

Chaurasia A, Culurciello E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. *IEEE Visual Communications and Image Processing (VCIP). IEEE*, 1-4.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv*, 1706-05587.

Cheng G. L, Wang Y, Xu S. B., 2017. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*,55(6), 3322-3337.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M., 2022. Swin-UNet: Unet-like pure Transformer for medical image segmentation. *In European conference on computer vision,* 205-218.

Carion Nicolas, Massa Francisco, Synnaeve Gabriel, 2020.End-to-end object detection with transformers. *European Conference on Computer Vision*, 213-229.

Chen L C, Papandreou G, Kokkinos I, 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.

Dosovitskiy Alexey,Beyer Lucas,Kolesnikov Alexander, 2021. An image is worth 16 × 16 words:transformers for image recognition at scale. *International Conference on Learning Representations*, 1-22.
Gao, X., Sun, X., Zhang, Y., Yan, M., Xu, G., Sun, H., ... & Fu, K., 2018. An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network. *IEEE Access*, 6, 39401-39414.

He, K., Zhang, X., Ren, S., & Sun, J., 2016. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Liu Ze,Lin Yu-tong,Cao Yue,et al., 2021. Swin transformer:hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*,10012-10022.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S., 2017. Feature pyramid networks for object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-2125.

Mendes C C T, Frémont V, Wolf D F., 2016. Exploiting fully convolutional neural networks for fast road detection. *IEEE International Conference on Robotics and Automation. IEEE*, 3174-3179.

Mnih V, Hinton G E., 2010. Learning to detect roads in high-resolution aerial images. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Proceedings, Part VI 11. Springer Berlin Heidelberg*, 210-223.

Parmar Niki,Vaswani Ashish,Uszkoreit Jakob,et al.Image transformer, 2018. *International Conference on Machine Learning*, 4055-4064.

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V., 2021. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9, 82031-82057.

Yuan, Q., Shafri, H. Z. M., Alias, A. H., & Hashim, S. J. B., 2021. Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and LiDAR data. *Remote Sensing*, 13(13), 2473.

Yuan, Q., Ang, Y., & Shafri, H. Z. M., 2021. Hyperspectral image classification using residual 2d and 3d convolutional neural network joint attention model. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44, 187-193.

Yuan, Q., & Mohd Shafri, H. Z., 2022. Multi-Modal Feature Fusion Network with Adaptive Center Point Detector for Building Instance Extraction. *Remote Sensing*, 14(19), 4920.

Yuan, Q., & Wang, N., 2022. Buildings Change Detection Using High-Resolution Remote Sensing Images with Self-Attention Knowledge Distillation and Multiscale Change-Aware Module. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *46*, 225-231.

Zhu, Q., Zhang, Y., Wang, L., Zhong, Y., Guan, Q., Lu, X., ... & Li, D., 2021. A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175, 353-365.

Zhong Z, Li J, Cui W, 2016. Fully convolutional networks for building and road extraction: Preliminary results. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE*, 1591-1594.

Zhou L C, Zhang C, Wu M., 2018. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE*, 192-1924.