# Fourier-Deformable Convolution Network for Road Segmentation From Remote Sensing Images

Huajun Liu, *Member, IEEE*, Xinyu Zhou, Cailing Wang, Suting Chen, and Hui Kong, *Member, IEEE*

***Abstract*— Road segmentation from remote sensing images is a challenging task in capturing weak, long, and irregular road features due to the limited connectivity-preserving modeling capability. In this work, we proposed a U-shaped Fourier-deformable convolution network (FDNet) for road segmentation, which integrates the merits of deformable convolutions (DCs) and Fourier convolutions compactly. Specifically, a saliency-aware DC (SD-Conv) layer is proposed for tracing salient road features based on an iterative dynamic offset learning mechanism to grasp extremely tender and weak road objects. Meanwhile, a lightweight global feature extracting module based on spectral convolutions, namely, the adaptive Fourier convolution (AF-Conv) layer, is adopted to learn long-range dependency to extract long and continuous road structures. The proposed SD-Conv layer worked in parallel with the AF-Conv layer to construct a basic and compact block to build the U-shaped FDNet model for road segmentation. Furthermore, to maintain the continuity of road objects in complex road conditions, we introduced a topology-oriented loss function based on the Hausdorff distance (HD) on the persistence diagram (PD) of segmented results, and further combined with softDice loss components for fully supervised training. Our FDNet has been trained and evaluated on two benchmarks, and experimental results show that FDNet achieved state-of-the-art (SOTA) performance. Specifically, it achieved 80.34% on accuracy, 88.42% on precision, and 84.70% on mean intersection over union (mIoU), respectively, on the Massachusetts dataset, and achieved 99.05% on accuracy, 89.21% on precision, 88.61% on recall, and 81.37% on mIoU, respectively, on the DeepGlobe dataset, outperforming most previous methods on both datasets. Codes are available at: https://github.com/zhoucharming/FDNet.**

***Index Terms*— Fourier-deformable convolution network (FDNet), remote sensing images, road segmentation, saliency-aware deformable convolution (SD-Conv).**

Received 8 May 2024; revised 26 August 2024; accepted 25 September 2024. Date of publication 8 October 2024; date of current version 18 October 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62272234 and in part by the Fundo para o Desenvolvimento das Ciências e da Tecnologia of Macau (FDCT) under Grant 0067/2023/AFJ. *(Corresponding authors: Huajun Liu; Hui Kong.)*

Huajun Liu and Xinyu Zhou are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: liuhj@njust.edu.cn).

Cailing Wang is with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210000, China.

Suting Chen is with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Hui Kong is with the Department of Computer and Information Science (CIS), University of Macau (UM), Macau, China (e-mail: huikong@um.edu.mo).

Digital Object Identifier 10.1109/TGRS.2024.3476087

## I. Introduction

ROAD segmentation from remote sensing images is a crucial task for many applications, including urban planning, intelligent transportation systems, autonomous driving, vehicle navigation, road monitoring, and emergency management [1], [2]. Nevertheless, road segmentation from remote sensing images still poses serious challenges, including irregular road networks, complex backgrounds, and occlusions caused by trees and buildings [3].

In earlier works, researchers extensively studied how to extract road objects from high-resolution remote sensing images [4], [5], [6] by traditional image analysis approaches, which depend on too much expert knowledge and manual parameter tuning. Those approaches usually rely on structural features [7], i.e., texture, contrast, and shape, the difference from morphology [8] or spectral features [9] between road objects and background. Those traditional approaches demonstrate favorable results on some well-established scenarios but often require more prior knowledge and much expert experience, which would fail to generalize to more general cases [10], [11].

Deep neural networks have boosted road segmentation performance to a higher level than traditional methods. The classical convolution neural networks (CNNs), such as SegNet [12], UNet [13], and their variants, i.e., D-LinkNet [14] and ConDinet++ [15] have achieved promising results benefiting from their merits of parameter sharing and translation invariant, and powerful capability of crisp feature capturing for road segmentation tasks. For instance, the classical fully convolution network (FCN) [16] is the earlier model for road segmentation, which uses fully connected layers to replace de-convolutions in the decoder to achieve pixel-level prediction and can successfully grasp continuous road features [17]. Jie et al. [18] proposed the MECA-Net, which leverages convolution kernels of different sizes ($1 \times 1$, $3 \times 3$, and $5 \times 5$) to aggregate multiscale road features and further incorporates the channel attention module and strip pooling module in parallel fashion to capture long-range contextual information from both the channel and spatial dimensions. Similarly, the GMR-Net, proposed by Zhang et al. [19], utilizes the multiscale dilated convolutions and global attention mechanism to capture different scales of road features, which can effectively filter out ambiguous features by feature aggregating mechanism. Since the deformable convolution network (DCN) [20]

has been introduced by augmenting the spatial sampling locations with additional offsets, deformable convolutions (DCs) have been proposed for semantic segmentation tasks. For instance, the restricted DC (RDC) [21] is proposed for road scene semantic segmentation through a multitask learning architecture. In RoadFormer, Jiang et al. [22] developed a pyramidal deformable transformer architecture to extract road networks with remote sensing images by a multicontext patch embedding scheme. Furthermore, Dai et al. [23] proposed a road-augmented deformable attention network (RADANet) to learn the long-range dependencies for specific road pixels by the prior knowledge of the road shape and the progress in DCs and achieved satisfactory results. However, the inherent locality of these convolutional operations still limits the CNNs on grasping long and continuous road objects in remote sensing images, and the square sampling pattern of classical CNNs is not optimally suitable for their linear structures of road structures. Moreover, the classical DCs with fixed patterns cannot capture the diverse shapes of road structures effectively.

The transformer model and its variants [22], [24], [25], [26] provide a comprehensive framework for simultaneous global and local feature extraction for vision-related tasks and have achieved satisfactory results for road segmentation. Recently, transformer architectures [22], [27] have attracted widespread interest in the remote sensing image segmentation tasks because of the self-attention mechanism that can capture long-range features. For instance, Zhang et al. [27] proposed DCS-TransUperNet as a new framework based on transformer encoder, which mainly borrows the merits of CSwin transformer [26] and UperNet [28] architecture, to alleviate the inherent induction bias of CNNs for automatic remote sensing image segmentation. Similarly, Jiang et al. [22] built a road segmentation network that utilized the Swin transformer [24] as its backbone. One primary challenge of applying self-attention models on vision tasks is the considerable computational complexity that grows quadratically as the number of tokens increases. Furthermore, another study proposed by Wang et al. [29] analyzed the spectrum of features formed by using self-attention and pointed out that it behaves like low-pass filters that would suppress high-frequency signals and smoothen detailed feature maps, which means that these blocks are undesirable for tender objects segmentation from remote sensing images.

Although classical CNN and transformers-dominating deep neural networks have achieved amazing results on remote sensing image analysis, there still exist some challenges for reliable road segmentation based on these classical methods, including 1) the roads can vary in size, shape, texture, and surroundings, fixed convolutional patterns making segmentation challenging; 2) roads often blend with surrounding features like buildings, vegetation, and shadows, making them hard to be distinguished on the spatial domain; 3) remote sensing images may suffer from issues like noise, blur, or low contrast and the road objects usually are tender and weak, affecting segmentation accuracy; and 4) roads hold geometric similarity to rivers and gullies so that it is prone to misclassification.

Beyond transformer, the spectral convolution theorem [30] provides another paradigm to design novel neural architectures to capture global features by locally operating in the spectral domain. Moreover, road areas in remote sensing images have obvious low-frequency characteristics and show continuity. The areas beside the road have varying-frequency domain characteristics (for example, much higher frequencies). Therefore, road segmentation based on spectral-domain learning is a more efficient and effective way. Furthermore, these Fourier-based works have shown fine-grained modulation learning with fractional convolution [31], Gabor filter [32], or other filters [33] on spectral-domain-based complex feature space has powerful capability to capture richer semantic road features in remote sensing images, and can enhance the discrimination between anomalies and background. Several operators on the Fourier domain, such as fast Fourier convolution (FFC) [30], global filter networks (GFNets) [34], and adaptive Fourier neural operator (AFNO) [35], have shown more flexibility and efficiency on diverse vision-related tasks. Moreover, learning on the frequency domain can be achieved by transforming images from the spatial domain to the spectral domain by utilizing frequency filters to attenuate or enhance the high-frequency components and low-frequency components. This enables the enhancement of the differences between the tender boundaries and backgrounds in the remote sensing image and reduces interference of occlusion of buildings. Specifically, for road segmentation tasks on remote sensing images, low-pass filters can be applied to suppress the changes in the low-frequency information and to grasp the small-sized objects, thereby narrowing the intraclass difference between rivers and gullies. On the other hand, high-pass filters can strengthen the high-frequency information and expand the interclass difference of similar objects. Overall, frequency-domain operators [33] provide a flexible feature representation method to detect weak and continuous road features from remote sensing images.

Inspired by the DCNs [20] and frequency-domain operators [33], we further investigate learnable and dynamic offsets of novel DCNs to capture the diverse branches of tender and weak road structures and combine Fourier-domain operators to grasp long and continuous road trunk objects simultaneously. Specifically, we propose a novel neural architecture named Fourier-DCN (FDNet), where the DC kernels aim to adaptively focus on branches of tender and weak road structures, and the spectral-domain operators aim to grasp long and continuous road features. In addition, based on the persistent homology (PH) constraint theory [36], we rebuilt a topology-oriented loss function, which combined the softDice loss [37] to train the model to maintain the continuity of road objects. The main contributions of this work can be summarized as follows.

1) We proposed a novel architecture for road segmentation in remote sensing images, named U-shaped FDNet, which is powerful in capturing long and continuous, weak and tender road features by mixing adaptive Fourier convolutions (AF-Conv) and saliency-aware DC (SD-Conv) compactly.
2) We proposed a new SD-Conv, where the DC is based on a dynamic offset learning mechanism, and the offset can be traced along salient semantic features with an

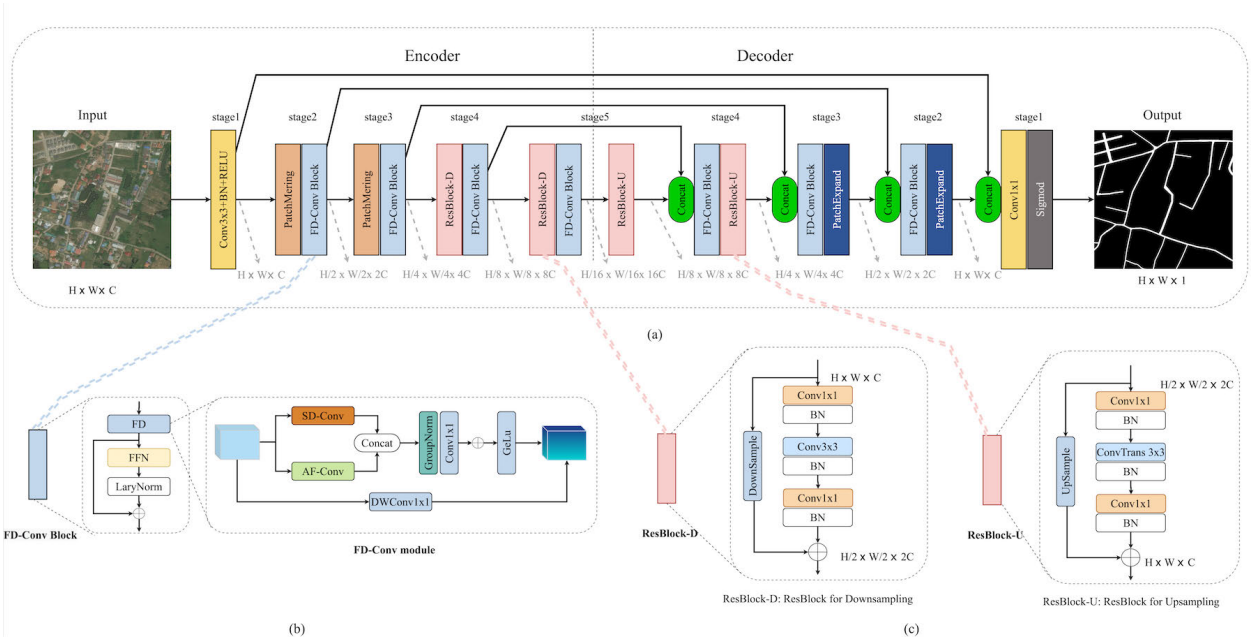


Fig. 1. (a) Overall structure of the proposed FDNet. (b) Components of the FD-Conv module, where SD-Conv stands for saliency-aware DC, and AF-Conv stands for adaptive frequency convolution [38]. (c) Components of ResBlock-D and ResBlock-U.

iterative strategy to grasp extremely tender and versatile road features.

3) We proposed a topology-oriented loss function based on Hausdorff distance (HD) on the persistence diagram (PD) and combined it with the softDice loss to mitigate the segmentation problem on nonsmooth and narrow roads, further improving the continuity of road segmentation.
4) The proposed FDNet achieved state-of-the-art (SOTA) performance on two public benchmarks including the Massachusetts and DeepGlobe datasets.

The rest of this article is organized as follows. The methodology of the proposal is introduced in detail in Section II. Then, the comprehensive information about the road segmentation dataset, data augmentation, and performance metrics is described in Section III. The experimental results and ablation studies are stated in Section IV. Section V presents the conclusion of this work.

## II. Methodology

A new Fourier mixing deformable CNN is suggested for road segmentation from remote sensing images. The proposed method consisting of a basic feature representation block named SD-Conv, a deep learning model named FDNet following the original UNet pipeline with better performance, and a new topology-oriented loss function, is used for binary road prediction from remote sensing images. The following explains the architecture of the FDNet model and the topology-oriented loss function.

### A. Overall Architecture

The architecture of the proposed model is depicted in Fig. 1. The basic architecture of FDNet follows the classical UNet [13] pipeline, including a symmetric encoder and decoder part with stage-wise skip-connection for shallow to deep feature fusion. Initially, the encoder starts with a stem block consisting of a $3 \times 3$ convolution layer, batch normalization (BN), and a rectified linear unit (RELU) layer. It aims to extract high-dimensional features from the original remote sensing RGB image. The encoder features are fed into the subsequent feature encoding modules, mainly including our proposed FD-Conv block and the ResBlock in classical ResNet [39] stage by stage. The patch-merging layer further refines these features. The FD-Conv block for long and continuous feature extraction in the shallow stage (i.e., stages 2 and 3) is shown in Fig. 1(b). The patch-merging layer includes a single linear layer used for down-sampling. While in the deeper stages (i.e., stages 4 and 5), the ResBlock-D defined in the ResNet [39] [see Fig. 1(c)] is employed for fined and tender road feature capturing, which is essentially a ResBlock layer with smaller kernels, such as $1 \times 1$, $3 \times 3$, and closely followed by BN layers for regularization simultaneously.

In the decoder part, five corresponding stages are symmetrical to the encoder parts, aiming to reconstruct predicted semantic information from the high-dimensional feature space. In stage 5, the features output by the FD-Conv block and ResBlock-D will be up-sampled by a closely following ResBlock-U [see Fig. 1(c)]. In stages 4, 3, and 2, the features obtained from the encoder and corresponding features derived from the preceding stage in the decoder are concatenated to combine the shallow and deep features. Subsequently, the concatenated features are fed into our proposed FD-Conv blocks, enabling the features to incorporate crisp road features and diverse trunk road ones. Subsequently, these features are up-sampled by a patch-expanding block consisting of a linear layer to the shallow decoder stages. Finally, the reconstructed feature map with the original size can be obtained by a

convolution layer with $1 \times 1$ kernel and be further used to predict the probability of the road objects and background by a Sigmoid function.

It should be mentioned that we employ two different resampling methods in the encoder and decoder parts. The rationale is that shallow features have high resolution on spatial dimension, where we utilize linear transformation of patch-merging and patch-extending for up-sampling and down-sampling. In deeper stages, where the spatial resolution is lower, we use convolution operations as well as transpose convolution for up-sampling and down-sampling, respectively, and aim to maintain the detailed information in the low-resolution feature maps.

In the FDNet, the proposed FD-Conv block consists of SD-Conv and AF-Conv layers. It is the core block of the FDNet and is designed parallel for fine and weak road-like objects extracting and long and continuous road-like structure grasping. In our FDNet, utilizing convolutional down-sampling and up-sampling strategy in the deeper stage instead of linear layers in the original model significantly mitigates the serious issue of feature degenerating, effectively learns to align features in decoder and encoder parts, and maximizes the interaction of semantic features across channels simultaneously.

### B. Fourier-DC Module

The internal structure of the proposed FD-Conv block, which integrates the merits of DCs and Fourier neural operators, is illustrated in Fig. 1(b). Specifically, it consists of a novel SD-Conv layer to capture fine and weak road objects and a lightweight global feature extraction module, called the AF-Conv layer, to extract long and continuous road structures by modulation learning mechanism with point-wise convolutions on frequency-domain.

The SD-Conv layer and the AF-Conv layer simultaneously acquire local details and global structural features, and those features are concatenated along the channel dimension. These separate features from the spatial and frequency domains are combined with a parallel mode and the "GroupNorm" is utilized to normalize the channels within the submodules independently, which will help the model address the issue of internal corporate bias [40]. The two parts of the features are then combined using a convolution layer, whose kernel size is $1 \times 1$, to project the output channels to be expected.

### C. Saliency-Aware DC

DC [21], deformable transformer [22], and deformable attention network [23] have shown great potential for road segmentation with remote sensing images meriting from their advantages on the progress in deformable features capture or the long-range modeling capabilities. In the classical DCs, it usually integrates learnable deformation offsets imposed on the convolutional kernel to capture shape-aware features. For instance, the constrained DC in DSConv [41] uses two individual and orthogonal $1 \times 3$ and $3 \times 1$ convolution kernels in the $x$- and $y$-axial directions sequentially, and then both features are concatenated to obtain a fused feature. Inspired by the soul of DCNs [20] and DSConv [41], we rethink the DCs and extend it to the dynamic offset learning problem, where the offset can be traced along salient semantic features with an iterative strategy at four directions from kernel's center to peripheries to grasp extremely tender and versatile road features.

Let us take a $C \times C$ convolutional kernel as an instance. Given a standard 2-D convolution [39], its $C \times C$ convolution kernel $\mathbf{K}$ can be formulated as

$$\mathbf{K} = \left\{\mathbf{K}_{(i,j)}, i, j \in [-C/2, C/2]\right\} \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{C \times C \times W}$ is the convolution kernel, $\mathbf{K}_{(i,j)} \in \mathbb{R}^{W}$ is the weights vector with coordinate offsets of $(i, j)$, and $W$ is the dimension of weight and bias (default as 2), the sampling pattern of each convolution kernel (short for $P_1$) can be formulated as

$$P_1(\mathbf{K}^{(h,w)}) = \{(x_{i,j}, y_{i,j}), i, j \in [-C/2, C/2]\} \tag{2}$$

where $\mathbf{K}^{(h,w)}$ is a convolution kernel with $(h, w)$ central coordinate, and $\mathbf{K}^{(h,w)}_{(i,j)}$ is a weights vector with $(h, w)$ central coordinate and $(i, j)$ coordinate offset, then $(x_{i,j}, y_{i,j})$ records the vector's absolute coordinate, which means $x_{i,j} = h+i$ and $y_{i,j} = w + j$.

Moreover, in a DC [20], the deformation offset $\Delta$ extends the fixed sampling pattern to the variable sampling pattern, where the sampling pattern (short for $P_2$) in a DC can be expressed as

$$P_2(\mathbf{K}^{(h,w)}) = \{(x_{i,j} + \Delta x, y_{i,j} + \Delta y), i, j \in [-C/2, C/2]\} \tag{3}$$

where $\Delta x$ and $\Delta y$ represent the coordinate offsets to be learned, and $\Delta x, \Delta y \in [-H, H]$, where $H$ represents the size of feature map.

Furthermore, in our SD-Conv layer, we impose the dynamic offset learning strategy on the classical DC. The center in a convolution kernel is taken as the origin, and the dynamic offset will be constrained along four axial directions, i.e., up, down, left, and right. Dynamic sampling positions from center to peripheries will be sequentially traced along salient features, enabling the kernel to capture long and continuous road features. Specifically, given a convolution kernel $\mathbf{K}^{(h,w)} \in \mathbb{R}^{C \times C \times W}$, which is a matrix centered at $(h, w)$. Then, the sampling pattern of our SD-Conv kernel (short for $P_3$) can be defined as

$$P_3(\mathbf{K}^{(h,w)}) = \{P_u, P_d, P_l, P_r\} \tag{4}$$

where $P_u$, $P_d$, $P_l$, and $P_r$ represent the sampling coordinates of traced salient features along four axial directions, i.e., up, down, left, and right, respectively, and the iterative process can be defined in detail. And the sampling coordinates along "up" direction of the kernel's slice $\mathbf{K}^{(h,w)}_{(0,a)}$ follows on $\mathbf{K}^{(h,w)}_{(0,a-1)}$, is defined as a kernel's slice centered at $(h, w)$ and with coordinate offset of $(0, a)$, where $a \in [1, C/2]$, which can be defined as follows:

$$P_u\left(\mathbf{K}^{(h,w)}_{(0,a)}\right) = \left(x_{0,a} + \Sigma_w^{w+a}\Delta x, y_{0,a} + \Sigma_w^{w+a}\Delta y\right). \tag{5}$$

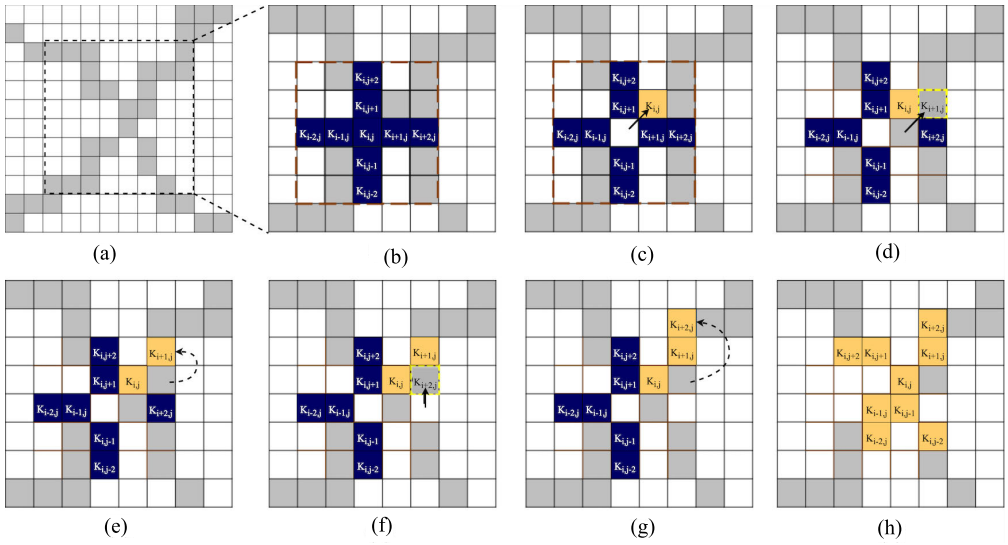


Fig. 2. Diagrams show the detailed process of saliency-aware deformation of the constrained kernel. In the first row, blue represents the initial state, and yellow represents the deformed state. (a) Feature map extracted at a certain stage, where the gray block is for the road and the light block is for the background. (b)–(g) Detailed process of cross-centered deformation of the constrained convolutional kernel. (h) Deformation process of the other four directions is the same as (b)–(g). It is worth mentioning that the deformation of the four directions occurs simultaneously.

The sampling coordinates along the "down" direction of the kernel's slice $\mathbf{K}^{(h,w)}_{(0,-a)}$ follows on $\mathbf{K}^{(h,w)}_{(0,-a+1)}$, which can be defined as follows:

$$P_d\left(\mathbf{K}^{(h,w)}_{(0,-a)}\right) = \left(x_{0,-a} + \Sigma^{w}_{w-a}\Delta x,\, y_{0,-a} + \Sigma^{w}_{w-a}\Delta y\right). \quad (6)$$

The sampling coordinates along the "left" direction of the kernel's slice $\mathbf{K}^{(h,w)}_{(-a-1,0)}$ follows on $\mathbf{K}^{(h,w)}_{(-a,0)}$, which can be defined as follows:

$$P_l\left(\mathbf{K}^{(h,w)}_{(-a,0)}\right) = \left(x_{-a,0} + \Sigma^{h}_{h-a}\Delta x,\, y_{-a,0} + \Sigma^{h}_{h-a}\Delta y\right). \quad (7)$$

The sampling coordinates along the "right" direction of the kernel's slice $\mathbf{K}^{(h,w)}_{(a+1,0)}$ follows on $\mathbf{K}^{(h,w)}_{(a,0)}$, which can be defined as follows:

$$P_r\left(\mathbf{K}^{(h,w)}_{(a,0)}\right) = \left(x_{a,0} + \Sigma^{h+a}_{h}\Delta x,\, y_{a,0} + \Sigma^{h+a}_{h}\Delta y\right) \quad (8)$$

where $\Delta x$ and $\Delta y$ are final coordinate offsets that need to be learned, and $\Delta x,\ \Delta y \in [-1, 1]$.

Based on the spatial iterative constraint of kernel sampling pattern, the convolution kernel and the dynamic offsets, where $\Delta x, \Delta y \in [-1, 1]$ can be learned to capture the deformation of road structure to trace the salient road features, which can be helpful to grasp extremely tender and continuous road structure [41]. The spatial iterative constraint is accomplished by an offset iterative learning strategy, which can steer the kernel to adaptively capture diverse road features from center to peripheries along four axial directions at each convolution operation. Moreover, our SD-Conv kernels integrate the merits of classical DCs and cross-constrained deformation strategy to simultaneously capture trunk and branch road features. The visualized sampling process of the SD-Conv kernel is shown in Fig. 2.

The learned sampling coordinates of all DC kernels are floating numbers, we need a bilinear interpolation to calculate the integer values of sampled points [20]. This process can be formulated as

$$\mathbf{P} = \Sigma_{\mathbf{P}'}\text{Bilinear}\left(\mathbf{P}', \mathbf{P}\right) \cdot \mathbf{P}' \quad (9)$$

where $\mathbf{P} \in \mathbb{R}^2$ denotes a fractional sampling location in (5)–(8), $\mathbf{P}' \in \mathbb{R}^2$ enumerates all probable integral spatial locations [41], and Bilinear$(\cdot, \cdot)$ is a bilinear interpolation operator. Subsequently, the feature maps can be refined by a resampled DC kernel through the sum of element-wise multiplication.

Our SD-Conv layer benefits from borrowing dynamic offset learning strategies of deformation convolution into a single convolutional kernel, and can dynamically capture diverse tender and continuous road features while preserving the locality and diversity of SD-Conv. In our FDNet kernel size of SD-Conv is 5. The algorithmic process of the SD-Conv is shown in Algorithm 1, where $k$[1] denotes the kernel size of a convolution.

### D. Lightweight Fourier Convolution

Inspired by the adaptive frequency filter neural network (AFFNet) [38], we use the spectral convolution theorem to build a computation-friendly global modeling method with lightweight local operations in the Fourier domain. The AFFNet employs the fast Fourier transform (FFT) with $\mathcal{O}(N \log N)$ computational complexity to model long-range dependency, which is lower than that of classical self-attention ($\mathcal{O}(N^2)$). In addition, rich spectral convolutions can provide refined global features by modulation learning on the spectral domain, and they have achieved amazing results on several vision-related tasks [33], [42].

We introduce the lightweight AF-Conv module based on the above work [38] to perform semantic frequency filtering

[1] $k$ is the abbreviation of kernel_size.

**Algorithm 1** Our Proposed SD-Conv

**Require:** *input* {% Input tensor of shape (N, H, W, C)}
**Ensure:** *output* {% Output tensor of shape (N, H, W, C)}
$N, H, W, C \leftarrow$ input.shape
$x \leftarrow input.permute()$ {% Rearrange input dimensions to N, C, H, W}
offset $\leftarrow offset(x)$ {% Calculate offset}
offset $\leftarrow reshapes$(offset, N, k, k, 2, H, W)
$c \leftarrow$ k//2
**for** $i \leftarrow 1$**to**$c$ **do**
  offset$[c, c+i] \leftarrow$ offset$[c, c+i-1]+$ offset$[c, c+i]$
  offset$[c, c-i] \leftarrow$ offset$[c, c-i+1]+$ offset$[c, c-i]$
  offset$[c+i, c] \leftarrow$ offset$[c+i-1, c]+$ offset$[c+i, c]$
  offset$[c-i, c] \leftarrow$ offset$[c-i+1, c]+$ offset$[c-i, c]$
**end for**
offset $\leftarrow reshape$(offset, N, k $*$ k $*$ 2, H, W)
mask $\leftarrow$ $mask(x)$ {% Calculate mask}
ref $\leftarrow get\_reference\_points$(input.shape, k, k)
grid $\leftarrow generate\_grids$(input.shape, k, k)
location $\leftarrow sum$(ref, grid, offset)
sampling_input $\leftarrow$ $grid\_sample$(input, location)
output $\leftarrow$ (sampling_input $*$ mask).$sum(-1).reshape$
**return** *output*

to grasp long and continuous road features. Given the input feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, the AF-Conv can be formulated as

$$\mathbf{X}' = \mathcal{F}^{-1}[\mathcal{M}(\mathcal{F}(\mathbf{X})) \odot \mathcal{F}(\mathbf{X})] \tag{10}$$

where $\mathcal{F}(\cdot)$ represents the FFT, which converts the feature maps $\mathbf{X}$ to features $\mathbf{X}_F \in \mathbb{C}^{H\times W\times C}$ in frequency-domain, and the adaptive mask $\mathcal{M}(\mathcal{F}(\mathbf{X})) \in \mathbb{C}^{H\times W\times C}$ is to be learned from the input feature $\mathbf{X}_F$ and possesses identical dimensions to $\mathbf{X}_F$. As depicted in Fig. 3, a lightweight convolution block consists of two sequential $1 \times 1$ depthwise convolution layers and closely follows a Relu activation function to generate a dynamic spectral mask to extract salient spectrum components. Subsequently, the salient spectral features can be refined by the Hadamard product $\odot$ operation on the mask and the feature map in the spectral domain. Finally, an inverse FFT (IFFT) $\mathcal{F}^{-1}(\cdot)$ converts the spectral features to the normal features. This operation enables a novel adaptive frequency filtering, which integrates lightweight operations, such as soft mask and lightweight convolutions, on the Fourier domain, and aims to grasp the refined global features to detect long and continuous road objects.

### *E. Loss Function*

In this section, we explore a topology-oriented loss function based on the topological data analysis [43] for road segmentation neural network. During training, introducing topological constraints to the loss function can help the model maintain the continuity of road features in complex road conditions. The PH [36] provides a path to compute topological features of a space at different resolutions. Particularly, as shown in Fig. 4, the evolution of 0-D and 1-D homological class when changing the scale parameter,

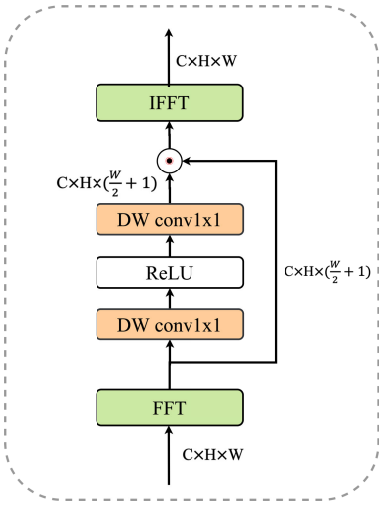


Fig. 3. Architecture of the adaptive frequency filter [38]. $\odot$ is the Hadamard product operation.

which is denoted by the radius of the light red circle around each data point [44]. The 0-D feature can describe connected components, which can depict linear road features well, and 1-D structures can describe connected components and ring structures [45], which means it can describe more complex topological relations in road segmentation tasks. Specifically, based on PH-generated PDs [46], we will define a topology-oriented loss function, as shown in Fig. 5, which integrates multiscale topological information and the distance definition between two clusters to impose the continuity constraint to the diverse road structure. In the PH theory [44], the birth time $b$ and death time $d$ of topological features are used to depict the evolution period. Such topological feature's distributions are summarized and plotted in the evolution space, which is a collection of points $(b, d)$ and called a PD [46]. Thus, the process of generating PD from binary segmentation results can be formulated as follows:

$$\Sigma_{(b,d)} = \mathrm{PDM}(\cdot) \tag{11}$$

where each point $(b, d)$ denotes the $k$th-PH class that appeared at time $b$ and disappeared at time $d$, and PDM$(\cdot)$ is a function to map the binary segmentation results to a PD. Specifically, in our road segmentation tasks, dimensions $k = 0, 1$ represent the connectivity components and ring structures, respectively [44].

The Hausdorff quantile [47] and HD [48] are classical metrics to measure the minor difference between the PDs of the annotated and the predicted results. However, the former usually excludes potential outlier predictions. However, in road segmentation tasks, outlier prediction is essential for extremely tender and weak road object detection. Therefore, in our loss function, the HD [48] is selected as a metric to measure the similarity of PDs between the ground truth and predicted results. The definition of HD on PDs is formulated as follows:

$$\begin{cases} \mathrm{dis}_{\mathrm{HD}}(\mathbf{P}_p, \mathbf{P}_g) = \max_{t\in\mathbf{P}_p} \min_{w\in\mathbf{P}_g} \|t-w\| \\ \mathrm{dis}_{\mathrm{HD}}(\mathbf{P}_g, \mathbf{P}_p) = \max_{w\in\mathbf{P}_g} \min_{t\in\mathbf{P}_p} \|w-t\| \\ \mathrm{dis}^*_{\mathrm{HD}} = \max\{\mathrm{dis}_{\mathrm{HD}}(\mathbf{P}_p, \mathbf{P}_g), \mathrm{dis}_{\mathrm{HD}}(\mathbf{P}_g, \mathbf{P}_p)\} \end{cases} \tag{12}$$

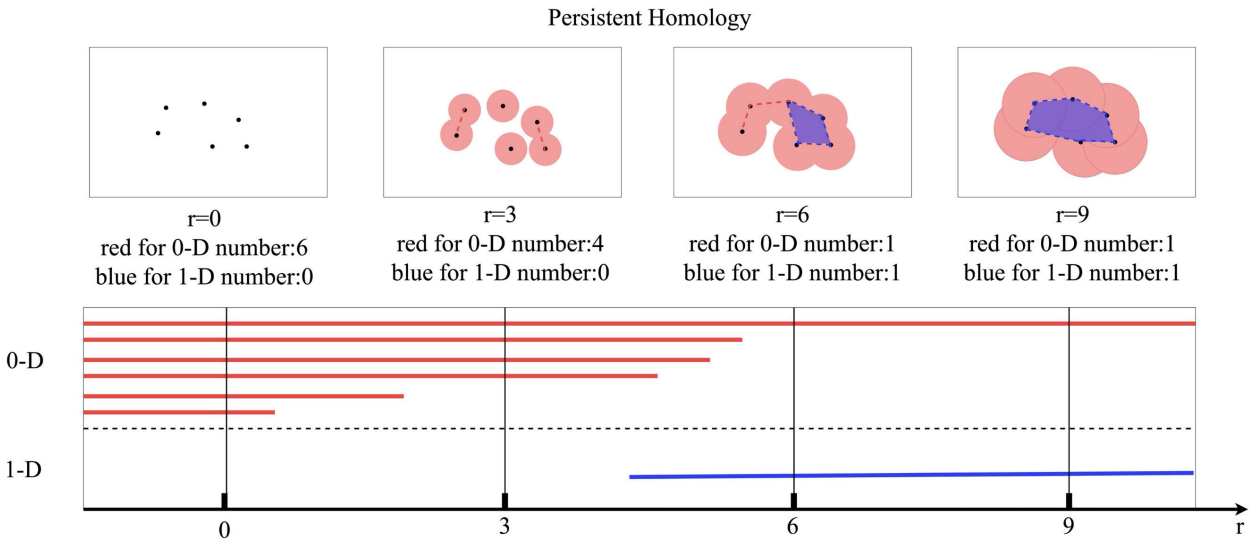


Fig. 4. Figure exhibits the process of PH [36]. The topological features of 0-D and 1-D homological classes are illustrated as red bars and blue bars, respectively, and $r$ is the radius adopted by points.

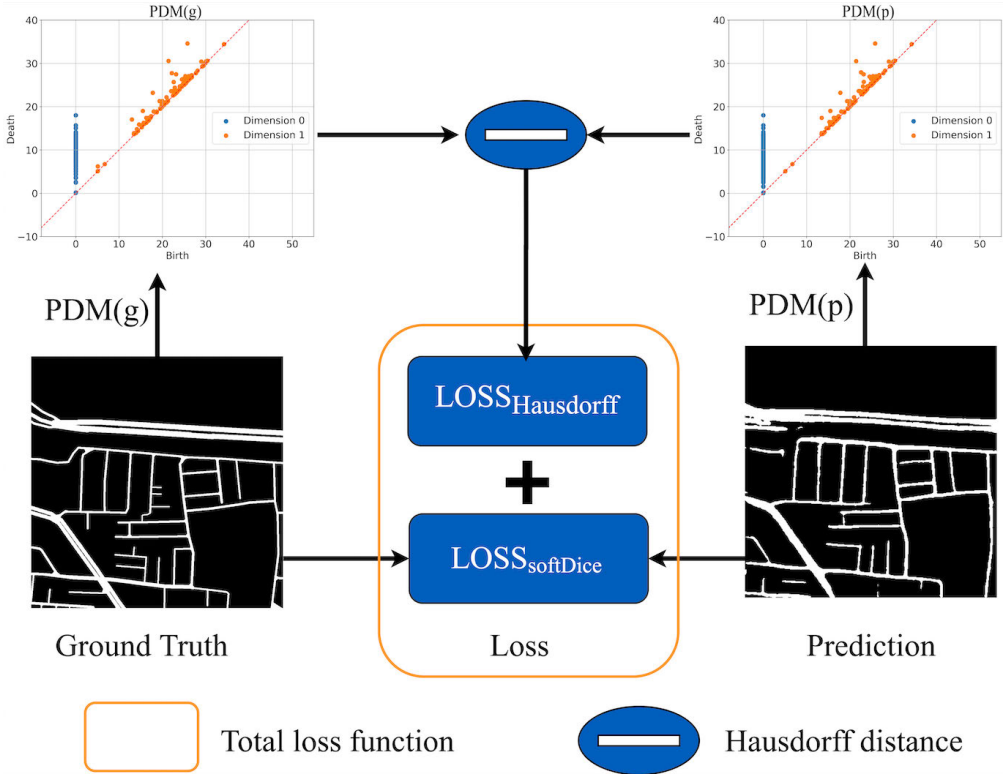


Fig. 5. Weighted loss function based on the HD of PDs and softDice.

where $t$ and $w$ are elements in a PD, respectively, $\mathrm{dis}_{\mathrm{HD}}(,)$ denotes the HD measuring of two PDs, and $\mathbf{P}_g \in \mathbb{R}^{H\times W\times 2}$ denoted the PD from annotated results, which is generated according to (11) and can be marked as PDM($g$), and $\mathbf{P}_p \in \mathbb{R}^{H\times W\times 2}$ denoted the PD from predicted results, which can be marked as PDM($p$), and $\mathrm{dis}^*_{\mathrm{HD}}$ denotes the final HD on PDs based on maximum selection mechanism.

It should be mentioned that HD is sensitive to outliers, which can help the model to capture the extremely tender and weak road features and improve the capability of connectivity-preserving, as shown in Fig. 6(c). Furthermore, the softDice [37] is a publicly accepted metric for vessel detection in medical image segmentation, which is a statistics-based loss function for unbalanced tasks [37]. Therefore, the final loss function is a weighted loss function that integrates the merits of the softDice and the HD on PDs. The final loss function is defined as follows:

$$L_{\mathrm{SH}} = (1 - \ \mathrm{softDice}\ ) + \lambda \sum_{n=0}^{N} \mathrm{dis}^*_{\mathrm{HD}} \tag{13}$$

$$\mathrm{softDice}\ = 2 \times \frac{M_P \cdot M_G}{M_P + M_G} \tag{14}$$

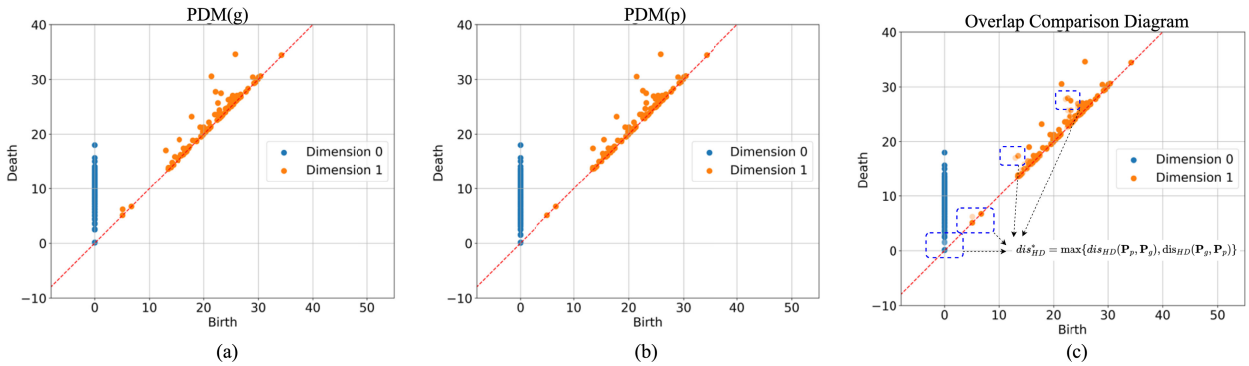


Fig. 6. Demonstration of the PD of binary segmentation results, where (a) denotes the PD of the ground truth, (b) denotes the PD of predicted results, and (c) shows an overlapping instance of them.

where $M_G$ is the ground truth mask and $M_P$ is the prediction mask, which are two binary masks [42]. A parameter $\lambda$ is used to determine the weight of the topological loss components, which is set to 0.5 in this case. The final loss $L_{\text{SH}}$ is a weighted one combining the softDice and the topological components.

## III. Datasets and Performance Metrics

### A. Datasets

*1) Datasets:* Two publicly available benchmarks, i.e., the Massachusetts Road Dataset and the DeepGlobe Road Dataset, are selected to train and evaluate our proposed FDNet model on road segmentation from remote sensing images. Below are the details of the two datasets.

*a) DeepGlobe dataset:* The DeepGlobe is a dataset prepared for the 2018 DeepGlobe Road Extraction Challenge. The dataset contains 6226 images with a resolution of 0.5 m and a size of 1024 × 1024 pixels. The images cover a variety of geographical settings in countries such as Thailand, India, and Indonesia, including roads in urban, suburban, and rural areas. Each annotated image is a three-channel binary image in PNG format, using (255, 255, 255) to indicate the road and (0, 0, 0) for presentation and background. Following the classical strategy in AFCNet [42], we divide the dataset into a training set of 4987 images and a testing set of 1246 images.

*b) Massachusetts dataset:* The Massachusetts Road dataset covers aerial imagery of the city of Boston and its surroundings. This dataset is divided into three parts: the training set contains 1108 images, the testing set contains 49 images, and the validation set contains 14 images, which is similar to the strategy in AFCNet [42]. All these images are with a resolution of 1500 × 1500 pixels.

*2) Data Augmentation:* The training set was conducted data augmentation by clipping, flipping, and rotation operations. To avoid generating the same images after rotating and flipping transformations, the rotating angles are set a 90°, 180°, and 270°, and the images are horizontally or vertically flipped to increase the amount of training set. Moreover, each image is cropped to a fixed size of 448 × 448 pixels with a crop stride of 448 pixels, and the remaining parts whose size is less than 448 will be conducted with the zero-filling method. During inference, the predicted results will be mosaiced to their original dimension.

### B. Performance Metrics

Road segmentation can be regarded as a binary classification problem, and the widely accepted metrics, such as precision, recall, $F$-score, and mean intersection over union (mIoU) are used for performance evaluation. The pixels in the segmented image are divided into true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Then, four metrics are calculated as follows.

The mIoU quantifies the overlap between the predicted segmentation region and the annotated ground truth region from a specific dataset, which is defined as follows:

$$\text{mIoU} = \frac{1}{2} \times \left( \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP} + \text{FN}} \right). \quad (15)$$

The precision is the percentage of correctly classified road segmentation results. It is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (16)$$

The accuracy is the ratio between the number of correctly classified pixels and the total number of pixels. It can be defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (17)$$

The recall refers to the proportion of all TP classes (TP + FN) that are judged to be positive classes (TP), which is calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (18)$$

Moreover, the efficiency analysis in terms of floating-point operations per second (FLOPs) and frames per second (FPS) is also used for computational complexity and inference speed evaluation, and all models are evaluated on a single Nvidia 3090 GPU.

TABLE I
CONFIGURATION OF LEARNING RATE COEFFICIENT

| Stage | Iterations | Power |
|---|---|---|
| 1 | 0K-4K | 0.9 |
| 2 | 4K-8K | 1.5 |
| 3 | 8K-12K | 2.5 |
| 4 | 12K-16K | 3.2 |

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

### A. Implementation Details

The proposed FDNet is trained with PyTorch 1.8 on a single NVIDIA RTX 3090 GPU. When training, we set the batch size to 4 and employ the stochastic gradient descent (SGD) optimizer to manage the learning rate. We adopt the polynomial decay strategy to update the learning rate stage-by-stage to decline the learning rate at the start of each stage. Specifically, the training process is divided into four phases, each with a different power decay coefficient which can adjust the learning rate at different stages of training to make the model approach the optimal solution, and the configuration is listed in the Table below.

The initial learning rate is set to $\lambda = 1e^{-2}$ and gradually reduced to $\lambda = 1e^{-7}$ by linear decay after 160 000 iterations (Table I). The stage-wise learning rate adjustment strategy is helpful for the model to converge steadily during training and achieve excellent segmentation performance.

### B. Comparison With the SOTA Methods

*1) Results on the DeepGlobe Dataset:* To comprehensively evaluate the segmentation performance of our model on the DeepGlobe dataset, we compared the visualized results of FDNet and several classical models, including U-Net [13], SwinU-Net [49], D-LinkNet [14], and Deep FR TransNet [50].

For those rural situations (e.g., the first to second rows in Fig. 7), where the vegetation or building occlusion usually exists and the road networks are relatively sparse, most baseline models are competent to segment the thick road parts. Nevertheless, there are many discontinuous segmentation results during the occlusion or variable road shapes exist. In contrast, our proposed FDNet shows amazing performance on these thin, narrow roads and probable occluded parts, and provides a more consistent capability to segment these rural road networks.

For those challenging scenarios with dense road networks (e.g., the third to the seventh rows in Fig. 7), most baseline models often fail to detect tender or crossing road objects because of blending with surrounding features and geometric similarity to rivers and gullies. In contrast, our FDNet demonstrates excellent segmentation performance in those challenging scenarios characterized by blending, similar, and dense road conditions.

Comparative experiments were conducted with other deep learning models, including classical convolutional models, such as FCN [16], SegNet [12], DeepLabv3 [52], U-Net [13], D-LinkNet [14], ConDinet++ [15], and MECA-Net [18], classical transformer model, such as TransU-Net [25] and SwinU-Net [49]. In addition, some recent advancements in the field, including SC-RoadDeepNet [54], AFU-Net [33], and CoAnet [53]. The quantitative performance comparison results are listed in Table II. Our FDNet achieved the best performance with 99.05% on accuracy, 89.21% on precision, 88.61% on recall, and 81.37% on mIoU metric. Specifically, the FDNet exceeded the recent convolutional models, such as MECA-Net [18] and obtained the mIoU gains of 16.22%, and the FDNet outperformed the classical transformer models, such as SwinU-Net [49] by 19.94% on mIoU, and it also obtained 3.11% mIoU gains compared with the recent Fourier-domain methods, such as AFCNet [42]. It can be concluded from Table II that our proposed FDNet outperformed most previous methods on the mentioned metrics, furthermore, FDNet can achieve satisfactory results only consuming moderate FLOPs.

*2) Results on the Massachusetts Dataset:* Scenarios in the Massachusetts dataset are more challenging because of its thin and dense road shapes, low contrast with surroundings, and probable occlusions with buildings and shadows. Moreover, the resolution in the Massachusetts road dataset is much lower, measuring at 1.5 m/pixel, but the resolution in the DeepGlobe dataset is 0.5 m/pixel. Consequently, the low resolution makes road shapes thinner in the Massachusetts road images, which requires the model to distinguish more accurately between these tender and thin road features.

The road segmentation results from classical models, such as U-Net [13], SwinU-Net [49], D-LinkNet [14], AFU-Net [33], and our proposed FDNet, were visualized for the selected scenarios. By analyzing the segmentation results in Fig. 8, it is clear that the SwinU-Net model, which benefits from the self-attention mechanism, can successfully detect the most detailed structure of the road network. However, it also has limitations in capturing the local structure of thin road objects and performs poorly on accurately detecting those ring structures in the Massachusetts road images, even though it is competent to preserve the overall road structure. In complex road intersection scenarios, most previous methods usually encounter difficulties in capturing the continuous structure.

By introducing the SD-Conv mechanism, our proposed FDNet exceeds most previous methods in accurately capturing

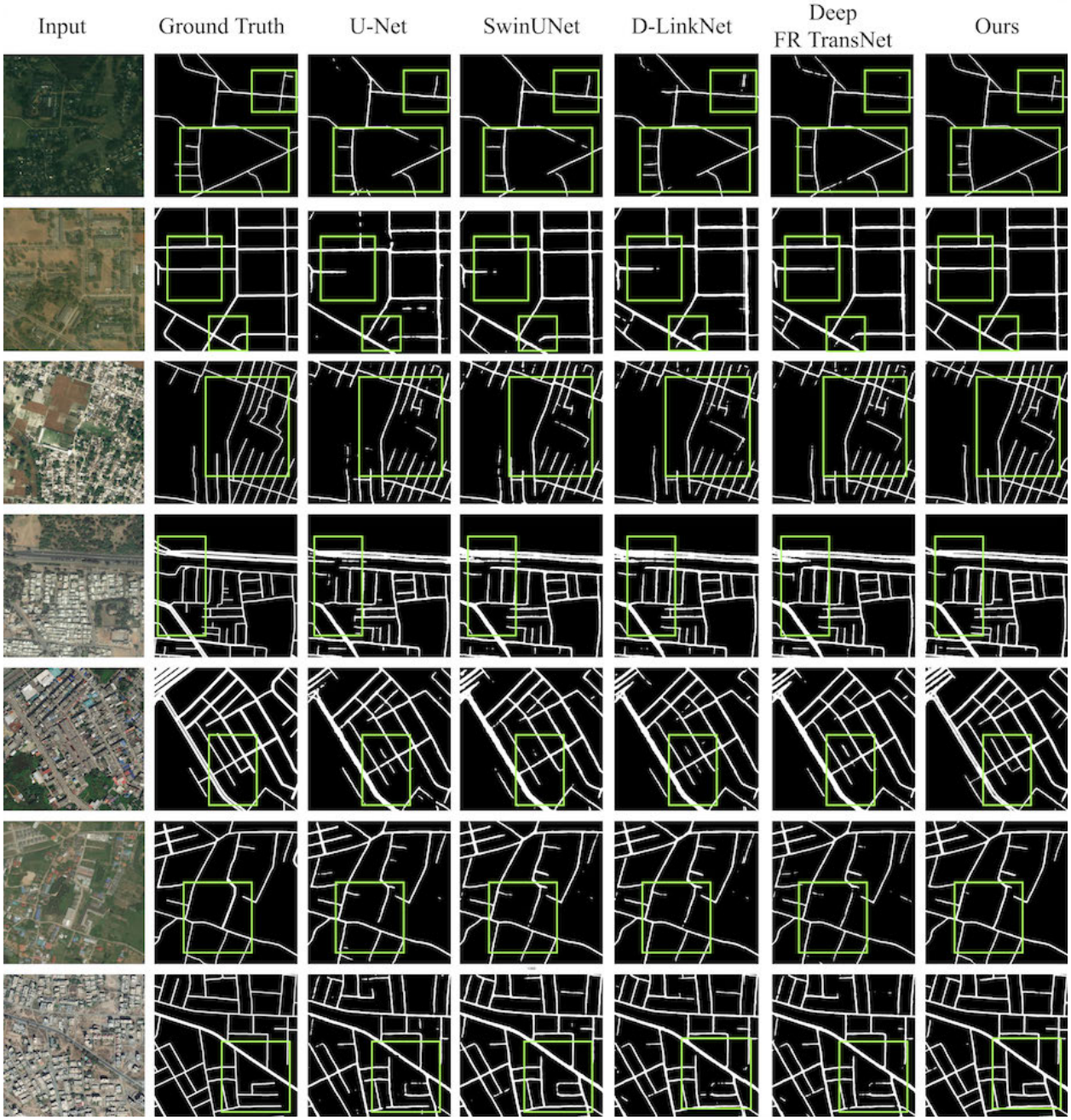


Fig. 7. Visualized results of our proposed FDNet and other deep learning models on the DeepGlobe dataset. The green box highlights areas where our model is superior to others.

thin and crossing road structures and further boosts the segmentation performance to a higher level under complex road conditions. In addition, our FDNet also introduces a topology-oriented loss function based on the PH theory, which guides the model to converge to more consistent results with the real scenarios, so that the model can detect complete road structures more accurately and distinguish those occlusions and similarities more consistently.

We compared FDNet with several baseline models, including FCN [16], SegNet [12], DeepLabv3 [52], TransU-Net [25], AttentionU-Net [51], and D-LinkNet [14]. The performance comparison results are listed in Table III. Inspired by recent advancements in the field, such as SC-RoadDeepNet [54], ConDinet++ [15], AFU-Net [33], and Seg-Road [55], we built the novel FDNet. And our FDNet achieved superior performance in terms of accuracy, precision, and mIoU, with values of 80.34%, 88.42%, and 84.70%, respectively. In addition, its recall is very close to the SOTA model (i.e., Seg-Road) [55]. Moreover, the Seg-Road utilizes the transformer as its backbone and conducts pretraining on the PASCAL VOC dataset [56], which would result in a significant increase in training costs. Nevertheless, our proposed FDNet outperformed the SOTA model (i.e., Seg-Road) [55] with 1.08% precision gains, and 0.81% mIoU gains, moreover, FDNet achieved 47 FPS under a single GPU, which is a competitive inference speed.

TABLE II
RESULTS ON DEEPGLOBE DATASET

| Model | Accuracy(%) | Precision(%) | Recall(%) | mIoU(%) | Params(M) | FLOPs(G) | FPS |
|---|---|---|---|---|---|---|---|
| FCN [16] | 92.05 | 63.71 | 61.54 | 42.45 | 22.70 | 25.34 | 49 |
| SegNet [12] | 91.54 | 62.03 | 60.21 | 41.85 | 29.43 | 123.04 | 59 |
| U-Net [13] | 97.12 | 82.28 | 71.20 | 61.02 | 31.04 | 156.89 | 47 |
| TransU-Net [25] | 93.21 | 64.45 | 62.34 | 46.10 | 67.62 | 99.51 | 50 |
| Attention-UNet [51] | 96.79 | 79.89 | 79.47 | 70.31 | 34.88 | 204.09 | 38 |
| SwinU-Net [49] | 97.01 | 82.35 | 71.54 | 61.43 | 40.14 | 69.75 | 45 |
| Deeplabv3 [52] | 95.87 | 65.83 | 64.48 | 48.31 | 68.11 | 206.48 | 32 |
| D-LinkNet [14] | 98.32 | 83.33 | 72.29 | 62.87 | 52.65 | 63.69 | 60 |
| MECA-Net [18] | - | 78.39 | 79.41 | 65.15 | - | - | - |
| Deep FR TransNet [50] | 98.70 | 87.30 | 81.15 | 72.44 | - | - | - |
| AFU-Net [33] | 98.81 | 87.86 | 81.70 | 74.14 | 110.04 | 191.34 | 14 |
| AFCNet [42] | 98.89 | 88.10 | 81.81 | 78.26 | 34.09 | 112.38 | 48 |
| CoANet [53] | - | 78.96 | 77.96 | 69.42 | - | - | - |
| CoANet-UB [15] | - | 86.54 | 85.91 | 80.69 | - | - | - |
| FDNet (Ours) | **99.05** | **89.21** | **88.61** | **81.37** | **36.85** | **125.41** | **47** |

### C. Results on the Wild Images

To demonstrate the generalization of our FDNet, we selected several satellite images and normal road images in the wild from Internet. Both dense and sparse road structures exist in these datasets, shown as Fig. 9. The result illustrates that the FDNet is capable of capturing both sparse and dense road networks effectively, demonstrating satisfactory segmentation results in various complex road structures.

### D. Ablation Studies

In this section, a series of ablation studies are conducted to investigate the components of FDNet, the inside SD-Conv layer, and further investigate the contribution of our proposed loss function.

*1) On Model Components:* To explore the contribution of basic components to the model performance, three optional sampling strategies, such as "All-ResBlock," "ResBlock + Linear," and "All-Linear," are first investigated in the FDNet. **All-ResBlock** uses ResBlock for convolution sampling operation between all four stages of the model. **ResBlock + Linear** (short for "hybrid strategy") is a hybrid sampling mode, where Patch-Merging is used for downsampling and Patch-Extand is used for upsampling in stages 2 and 3, while ResBlock is used for convolution sampling in stages 4 and 5. **All-Linear** uses linear transformations of Patch-Merging and Patch-Extand for up and down sampling, respectively.

It can be seen from Table IV that the hybrid sampling strategy shows significant performance improvement compared to the All-Linear sampling strategy. On the Massachusetts dataset, the hybrid strategy increases the mIoU from 83.35% to 84.7%, obtaining a gain of 1.35%. On the DeepGlobe dataset, the hybrid strategy increases the mIoU from 79.54% to 81.37%, obtaining a gain of 1.83%. However, in the All-ResBlock strategy, the number of model parameters increases to 54.66M, which is higher than the hybrid strategy's by 17.81M. Meanwhile, the FLOP of the All-ResBlock strategy reaches 218.64G, which is significantly higher than others. The linear transformation-based sampling method is cost-effective in resampling in the spatial dimension. However, the linear transformation will result in detailed feature loss and location ambiguity when up-sampling and down-sampling on spatial dimensions in the deep stage. The loss of details is significantly fatal for those tender and weak road scenarios, which requires the model to focus on fine features. Subsequently, the ResBlock was introduced to perform resampling by convolutions in the deeper stage, where the feature map is with lower resolution. It significantly reduces the issue of feature loss, obtains precise token localization, and enriches the semantic feature interaction across channels. Experiments demonstrate the effectiveness of the hybrid sampling mode, which serves as a better tradeoff between parameters consuming and segmentation results.

*2) On Module Structure:* The core FD-Conv block is built on the parallel combination fashion with the AF-Conv layer and our proposed SD-Conv layer, which serve as the basic and compact components.

The U-Net [13] was selected as the baseline model and several core blocks were introduced to replace the conventional convolution (CC). Ablation experiments on the Massachusetts and DeepGlobe datasets are designed to evaluate the individual contributions of each sublayer, namely, the SD-Conv layer and

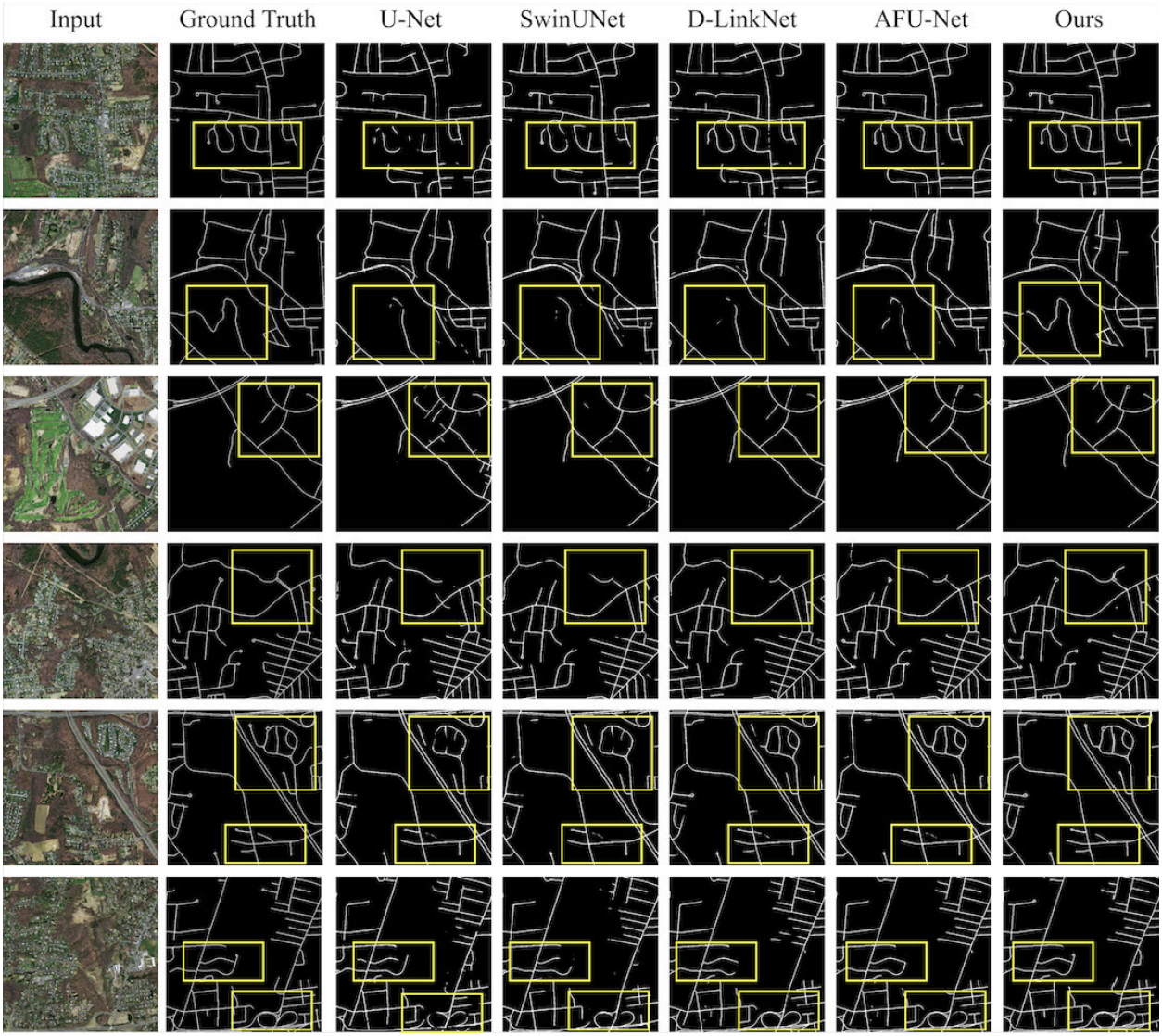


Fig. 8. Visualized results of FDNet and other deep learning models on the Massachusetts dataset. The yellow box highlights areas where our approach is superior to others.

the adaptive frequency filter (AF-Conv) layer, based on the performance metric of mIoU.

To evaluate the contribution of our proposed SD-Conv layer on extracting road features, we replaced the CC alone with the SD-Conv layer and the AF-Conv layer separately. From Table V, it is concluded that replacing the SD-Conv layer alone yields a greater improvement compared with replacing the AF-Conv layer alone. Specifically, the mIoU metric on the Massachusetts dataset increases from 80.33% to 83.72%, obtaining a gain of 3.39%. Similarly, on the DeepGlobe dataset, we got the mIoU metric from 72.77% to 79.74% and obtained a gain of 6.97%. To evaluate the efficacy of the fusion strategy involving the SD-Conv layer and AF-Conv layer, we integrate our proposed SD-Conv layer into the AF-Conv layer to build the complete FD-Conv module. It can be shown from Table V[2] that the SD-Conv layer boosted the model significantly. Specifically, upon introducing the SD-Conv layer into our model, we got the mIoU improvement from 80.33% to 84.70% on the Massachusetts dataset, obtaining a gain of 4.37%. Similarly, on the DeepGlobe dataset, the mIoU is enhanced from 72.77% to 81.37%, yielding a gain of 8.6%.

We added the AF-Conv layer on top of the SD-Conv layer, in which the complete FD-Conv block was formed, to assess the effectiveness of the AF-Conv layer. This increased to 0.94M parametric quantities and increased the FPS from 42 to 47, and improved the model performance in the Massachusetts and DeepGlobe datasets by 0.98% and 1.63%, respectively. The use of the lightweight AF-Conv layer in this study demonstrates how global Fourier domain features can enhance the model's perception of road features by providing information compensation for global structural features of the road.

The ablation experiment results demonstrate our proposed SD-Conv layer has the powerful capability to extract tender

[2]"Mass." and "Deep." are the abbreviations for the Massachusetts Road dataset and the DeepGlobe Road dataset, respectively.

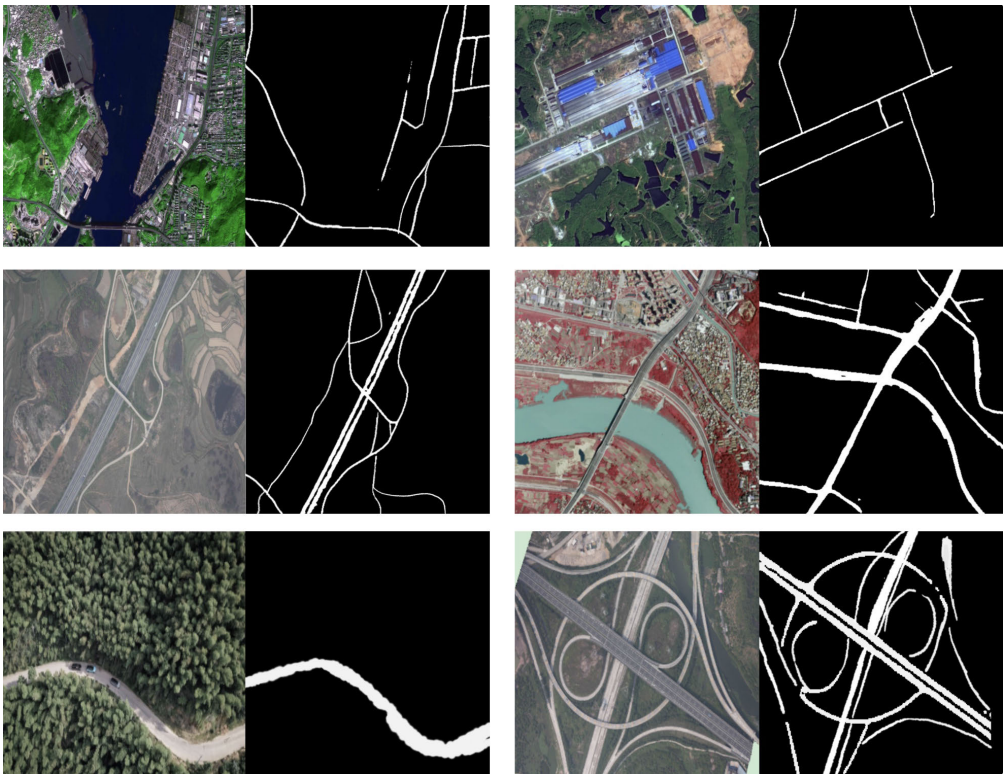

Fig. 9. Road images in the wild and the predicted result by our method.

TABLE III
RESULTS ON MASSACHUSETTS DATASET

| Model | Accuracy(%) | Precision(%) | Recall(%) | mIoU(%) | Params(M) | FLOPs(G) | FPS |
|---|---|---|---|---|---|---|---|
| FCN [16] | 74.61 | 60.43 | 68.89 | 73.20 | 22.70 | 25.34 | 49 |
| SegNet [12] | 75.23 | 64.41 | 69.80 | 74.81 | 29.43 | 123.04 | 59 |
| U-Net [13] | 78.04 | 63.23 | 70.32 | 76.39 | 31.04 | 156.89 | 47 |
| TransU-Net [25] | 78.10 | 68.35 | 71.23 | 77.89 | 67.62 | 99.51 | 50 |
| Attention-U-Net [51] | 76.31 | 71.32 | 73.41 | 78.13 | 34.88 | 204.09 | 38 |
| SwinU-Net [49] | 78.01 | 68.25 | 71.31 | 77.64 | 40.14 | 69.75 | 45 |
| Deeplabv3 [52] | 77.82 | 68.41 | 72.48 | 77.43 | 68.11 | 206.48 | 32 |
| D-LinkNet [14] | 77.82 | 67.16 | 71.73 | 77.23 | 52.65 | 63.69 | 60 |
| DCS-TransUperNet [27] | 77.94 | 69.52 | 73.48 | 77.74 | - | - | - |
| ConDinet++ [15] | 78.14 | 72.05 | 74.71 | 78.88 | - | - | - |
| AFU-Net [33] | 78.93 | 76.83 | 75.17 | 80.32 | 110.04 | 191.34 | 14 |
| SC-RoadDeepNet [54] | 76.53 | 76.97 | 74.89 | 82.69 | - | - | - |
| AFCNet [42] | 79.34 | 77.10 | 75.65 | 83.15 | 34.09 | 112.38 | 48 |
| Seg-Road [55] | - | 87.34 | **80.71** | 83.89 | 28.67 | - | 42 |
| FDNet (Ours) | **80.34** | **88.42** | 80.47 | **84.70** | **36.85** | **125.41** | **47** |

and diverse road features under complex road conditions. It also shows the effectiveness of the fusion strategy, which combines the SD-Conv layer (for local fine-grained features) and the AF-Conv layer (for global road structural features), to build our FD-Conv block. It is reasonable to conclude that the FD-Conv has great potential to be an efficient general-purpose layer for diverse road-structural representation learning, which can simultaneously capture

TABLE IV
ABLATION STUDY ON MODEL COMPONENTS

| Dataset | Sampling Strategy | mIoU (%) | Params(M) | FLOPs(G) | FPS |
|---|---|---|---|---|---|
| Mass. | All-ResBlock | 84.61 | 54.66 | 218.64 | 38 |
| | **ResBlock+Linear** | **84.70** | 36.85 | 125.41 | 47 |
| | All-Linear | 83.35 | 27.65 | 93.75 | 53 |
| Deep. | All-ResBlock | 81.34 | 54.66 | 218.64 | 38 |
| | **ResBlock+Linear** | **81.37** | 36.85 | 125.41 | 47 |
| | All-Linear | 79.54 | 27.65 | 93.75 | 53 |

TABLE V
ABLATION STUDY ON THE INSIDE OF FD-CONV BLOCK

| Dataset | Strategy | | | Metric | | | |
|---|---|---|---|---|---|---|---|
| | CC | SD-Conv | AF-Conv | mIoU(%) | Params(M) | FLOPs(G) | FPS |
| Mass. | ✓ | | | 76.88(baseline) | 31.04 | 107.52 | 45 |
| | | ✓ | | 83.72(+6.84) | 35.91 | 116.27 | 42 |
| | | | ✓ | 80.33(+3.45) | 28.09 | 87.63 | 50 |
| | | ✓ | ✓ | **84.70(+7.82)** | 36.85 | 125.41 | 47 |
| Deep. | ✓ | | | 61.02(baseline) | 31.04 | 107.52 | 45 |
| | | ✓ | | 79.74(+18.72) | 35.91 | 116.27 | 42 |
| | | | ✓ | 72.77(+11.75) | 28.09 | 87.63 | 50 |
| | | ✓ | ✓ | **81.37(+20.35)** | 36.85 | 125.41 | 47 |

TABLE VI
ABLATION STUDY ON DIFFERENT DEFORMATION STRATEGIES

| Dataset | Deformation strategies | | | Metric | | | |
|---|---|---|---|---|---|---|---|
| | DC [20] | DSC [41] | Ours | mIoU (%) | Params (M) | FLOPs(G) | FPS |
| Mass. | ✓ | | | 82.96(baseline) | 34.91 | 117.26 | 43 |
| | | ✓ | | 83.83(+0.87) | 44.71 | 158.39 | 41 |
| | | | ✓ | **84.70(+1.74)** | 36.85 | 125.41 | 47 |
| Deep. | ✓ | | | 77.68(baseline) | 34.91 | 117.26 | 43 |
| | | ✓ | | 79.74(+2.06) | 44.71 | 158.39 | 41 |
| | | | ✓ | **81.37(+1.63)** | 36.85 | 125.41 | 47 |

connectivity-preserved road features and extract global road structures in the Fourier domain.

*3) On Different Deformation Strategies:* By replacing our SD-Conv with classical DC [20] and DSConv [41], respectively, to evaluate the performance of our SD-Conv on both datasets, the results are listed in Table VI. The DSConv is good at extracting tender structural features using two individual 1 $\times$ 3 and 3 $\times$ 1 convolution kernels in the $x$- and $y$-axial directions sequentially. Whereas our SD-Conv integrates the merits of classical DCs and saliency-aware deformation strategy, which can capture more diverse road structures, and further reduce the parameters by 7.86M and reduce the FLOPs from 158.39G to 125.41G compared with the DSConv. Meanwhile, it improves the mIoU by by 1.74% and 1.63% on both datasets, respectively.

*4) On Loss Function:* To evaluate the contribution of our proposed loss function to the road segmentation task, we selected several well-established baseline models, namely,

TABLE VII
ABLATION STUDY ON LOSS FUNCTIONS. THE MIOU WAS CHOSEN AS THE VALIDATION METRIC

| Dataset | Model | $L_{softDice}$ | $L_{SH}$ (Ours) | **mIoU** (%) ↑ |
|---|---|---|---|---|
| Mass. | U-Net [13] | 76.39 | 76.88 | +0.49 |
| | SwinU-Net [49] | 77.43 | 78.09 | +0.66 |
| | Attention-U-Net [51] | 78.13 | 78.69 | +0.56 |
| | AFU-Net [33] | 80.32 | 80.84 | +0.52 |
| | FDNet | 83.94 | 84.70 | +0.73 |
| Deep. | U-Net [13] | 61.02 | 61.57 | +0.53 |
| | SwinU-Net [49] | 61.43 | 62.37 | +0.94 |
| | Attention-U-Net [51] | 70.31 | 70.63 | +0.32 |
| | AFU-Net [33] | 74.14 | 74.85 | +0.71 |
| | FDNet | 80.54 | 81.37 | +0.83 |

U-Net [13], SwinU-Net [49], Attention-U-Net [51], and AFU-Net [33], which also employ Fourier domain feature information for road feature extraction.

Several classical models and our proposal are trained on different loss functions. The validation experiment is conducted on both datasets, namely, Massachusetts and DeepGlobe. The mIoU was selected as the evaluation metric.

The results are presented in Table VII, as a consequence of our proposed loss function incorporating the HD method for PH theory. It is obvious from Table VII that $L_{\text{SH}}$ constrains the model topologically during the training process to give a better performance, that we observed a performance improvement of at least 0.3% to 0.8% in both road datasets. This improvement demonstrates the wide applicability of the $L_{\text{SH}}$ loss function across different models.

To evaluate the performance improvement achieved by the $L_{\text{SH}}$ loss function, we visualized the segmentation maps generated by FDNet training with $L_{\text{softDice}}$ and $L_{\text{SH}}$ loss function, respectively. Furthermore, we chose specific local regions for closer examination and analysis. As shown in Fig. 10, the yellow arrows highlight areas where there is a significant improvement in segmentation performance.

Fig. 10 indicates that the results generated by the $L_{\text{softDice}}$ training method are partially affected by breakage errors in locally complex road conditions. In addition, the segmentation results on wider roads are rough and not smooth, which is particularly evident in the DeepGlobe dataset. On the other hand, the results obtained from the $L_{\text{SH}}$ training method show a significant improvement in terms of continuity and smoothness. It should be noted that in some cases of parallel dual-roads, the segmentation results obtained using the $L_{\text{softDice}}$ loss function may not accurately segment the closer dual-roads. Instead, the model may interpret them as wider single roads. However, by introducing the $L_{\text{SH}}$ loss function with the topological constraint method, our model achieves good segmentation results on some of the adjacent dual roads as well, which benefits from the HD to outlier predictions. This sensitivity directs the model to prioritize pixels that contain outliers during training. Introducing the HD into the loss function effectively constrains the model to generate smoother and more continuous segmentation results, outperforming the ordinary $L_{\text{softDice}}$ loss function in terms of both segmentation accuracy and topological continuity.

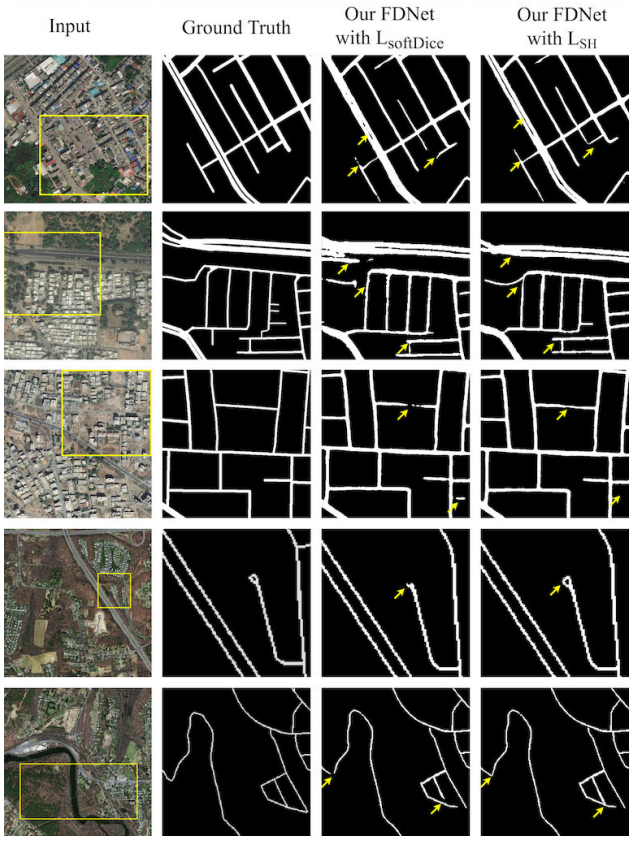


Fig. 10. Visualized results of two different loss functions. To evaluate the performance of our method, we selected some challenging regions in two datasets. In the following presentation, we present three rows of results from the DeepGlobe dataset and two rows from the Massachusetts Road dataset. From left to right, the raw image, the ground truth, and the results of FDNet are trained by $L_{\text{softDice}}$ and $L_{\text{SH}}$, respectively.

We also investigated the parameter $\lambda$ in our loss function. Our method achieved 82.73%, 84.22%, 83.81%, and 83.68% on the mIoU metric in the Massachusetts dataset when we used four different values of $\lambda = 0.2$, 0.5, 0.7, and 1, respectively. Therefore, the weight $\lambda$ in the final loss function should be set to 0.5.

## V. Conclusion

For the road segmentation task, we proposed a novel model called FDNet, which introduces a novel local–global feature representation pipeline for road segmentation tasks based on the SD-Conv and adaptive frequency filter. The SD-Conv architecture addressed the limitation of the original convolutional kernels on fixed patterns and local receptive fields and could effectively grasp diverse and dynamic road features. Meanwhile, we introduce an adaptive frequency filter layer to extract global structural features in a more efficient manner. Furthermore, we addressed the issues of inconsistent results in complex road conditions by a PH-based loss function.

In the future, we would like to conduct a broad and comprehensive digging on the FDNet on various tasks and explore building a more fundamental model for vessel-like structure modeling.

## References


[1] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.

[2] F. Yang, H. Wang, and Z. Jin, "A fusion network for road detection via spatial propagation and spatial transformation," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107141.

[3] Y. Wang et al., "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 4412612.

[4] Y. Bae, W.-H. Lee, Y. Choi, Y. W. Jeon, and J. B. Ra, "Automatic road extraction from remote sensing images based on a normalized second derivative map," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1858–1862, Sep. 2015.

[5] S. Valero, J. Chanussot, J. A. Benediktsson, H. Talbot, and B. Waske, "Advanced directional mathematical morphology for the detection of the road network in very high resolution remote sensing images," *Pattern Recognit. Lett.*, vol. 31, no. 10, pp. 1120–1127, Jul. 2010.

[6] 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 Observat. Remote Sens.*, vol. 5, no. 5, pp. 1538–1544, Oct. 2012.

[7] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for SAR and optical images using multiscale convolutional gradient features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[8] T. Géraud and J.-B. Mouret, "Fast road network extraction in satellite images using mathematical morphology and Markov random fields," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 16, pp. 1–12, Dec. 2004.

[9] X. Yang and G. Wen, "Road extraction from high-resolution remote sensing images using wavelet transform and Hough transform," in *Proc. 5th Int. Congr. Image Signal Process.*, Oct. 2012, pp. 1095–1099.

[10] I. Kahraman, I. R. Karas, and A. E. Akay, "Road extraction techniques from remote sensing images: A review," in *Proc. Int. Conf. Geomat. Geospatial Technol. Geospatial Disaster Risk Manag.*, vol. 42, 2018, pp. 339–342.

[11] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review," *Remote Sens.*, vol. 12, no. 9, p. 1444, 2020.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Springer, 2015, pp. 234–241.

[14] 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. CVPR Workshop*, 2018, pp. 182–186.

[15] K. Yang, J. Yi, A. Chen, J. Liu, and W. Chen, "ConDinet++: Full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] 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.

[17] Z. Zhong, J. Li, W. Cui, and H. Jiang, "Fully convolutional networks for building and road extraction: Preliminary results," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1591–1594.

[18] Y. Jie et al., "MECA-Net: A MultiScale feature encoding and long-range context-aware network for road extraction from remote sensing images," *Remote Sens.*, vol. 14, no. 21, p. 5342, Oct. 2022.

[19] Z. Zhang, X. Sun, and Y. Liu, "GMR-Net: Road-extraction network based on fusion of local and global information," *Remote Sens.*, vol. 14, no. 21, p. 5476, Oct. 2022.

[20] J. Dai et al., "Deformable convolutional networks," in *Proc. ICCV*, 2017, pp. 764–773.

[21] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang, "Restricted deformable convolution based road scene semantic segmentation using surround view cameras," 2018, *arXiv:1801.00708*.

[22] X. Jiang et al., "RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, Sep. 2022, Art. no. 102987.

[23] L. Dai, G. Zhang, and R. Zhang, "RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602213.

[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[25] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[26] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12114–12124.

[27] Z. Zhang, C. Miao, C. Liu, and Q. Tian, "DCS-TransUperNet: Road segmentation network based on CSwin transformer with dual resolution," *Appl. Sci.*, vol. 12, no. 7, p. 3511, Mar. 2022.

[28] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. ECCV*. Springer, 2018, pp. 418–434.

[29] P. Wang, W. Zheng, T. Chen, and Z. Wang, "Anti-oversmoothing in deep vision transformers via the Fourier domain analysis: From theory to practice," in *Proc. ICLR*, 2022.

[30] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. NIPS*, 2020, pp. 4479–4488.

[31] X. Zhao et al., "Fractional Fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2314–2326, Feb. 2024.

[32] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, "Fractional Gabor convolutional network for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5503818.

[33] J. Yang and H. Liu, "Modulation learning on Fourier-domain for road extraction from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[34] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. NIPS*, 2021, pp. 980–993.

[35] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, "Efficient token mixing for transformers via adaptive Fourier neural operators," in *Proc. ICLR*, 2022.

[36] A. Zomorodian and G. Carlsson, "Computing persistent homology," in *Proc. 20th Annu. Symp. Comput. Geometry*, Jun. 2004, pp. 347–356.

[37] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. 3rd Int. Workshop DLMIA*. Springer, 2017, pp. 240–248.
[38] Z. Huang, Z. Zhang, C. Lan, Z.-J. Zha, Y. Lu, and B. Guo, "Adaptive frequency filters as efficient global token mixers," 2023, *arXiv:2307.14008*.
[39] 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.
[40] Y. Sun et al., "Retentive network: A successor to transformer for large language models," 2023, *arXiv:2307.08621*.
[41] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6070–6079.
[42] H. Liu, C. Wang, J. Zhao, S. Chen, and H. Kong, "Adaptive Fourier convolution network for road segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5617214.
[43] M. E. Aktas, E. Akbas, and A. E. Fatmaoui, "Persistence homology of networks: Methods and applications," *Appl. Netw. Sci.*, vol. 4, no. 1, pp. 1–28, Dec. 2019.
[44] C.-C. Wong and C.-M. Vong, "Persistent homology based graph convolution network for fine-grained 3D shape segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7078–7087.
[45] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-preserving deep image segmentation," in *Proc. NeurIPS*, 2019.
[46] A. Patel, "Generalized persistence diagrams," *J. Appl. Comput. Topol.*, vol. 1, nos. 3–4, pp. 397–419, Jun. 2018.
[47] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, pp. 1–28, Dec. 2015.
[48] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
[49] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 205–218.
[50] Z. Ge, Y. Zhao, J. Wang, D. Wang, and Q. Si, "Deep feature-review transmit network of contour-enhanced road extraction from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[51] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[52] S. C. Yurtkulu, Y. H. Sahin, and G. Unal, "Semantic segmentation with extended DeepLabv3 architecture," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.
[53] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "CoANet: Connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021.
[54] A. Abdollahi, B. Pradhan, and A. Alamri, "SC-RoadDeepNet: A new shape and connectivity-preserving road extraction deep learning-based network from remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617815.
[55] J. Tao et al., "Seg-Road: A segmentation network for road extraction based on transformer and CNN with connectivity structures," *Remote Sens.*, vol. 15, no. 6, p. 1602, Mar. 2023.
[56] 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. 2015.

**Huajun Liu** (Member, IEEE) received the Ph.D. degree from the School of Computer Science, Nanjing University of Science and Technology, Nanjing, China, in 2007.

He worked as a Visiting Scholar and a Post-Doctoral Fellow at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 2018 to 2020. He is now with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include computer vision, information fusion, and deep learning.

**Xinyu Zhou** is currently pursuing the master's degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

His research interests include deep learning, computer vision, and image segmentation, especially in the fields of road segmentation based on remote sensing images.

**Cailing Wang** received the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology, Nanjing, China, in 2012.

She was a Visiting Scholar at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 2018 to 2019. She is now with the School of Automation and Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing. Her research interests include computer vision and deep learning.

**Suting Chen** is currently a Senior Professor with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China, and serves as a Doctoral Supervisor. Her professional fields mainly cover signal and information processing, computer vision, and many other research directions. In the industry, she holds more than 40 authorized patents and has published more than 70 academic papers at home and abroad.

**Hui Kong** (Member, IEEE) received the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, in 2007.

He is currently with the Department of Computer and Information Science (CIS), University of Macau (UM), Macau, China. Before that, he was affiliated with Nanjing University of Science and Technology (NJUST), Nanjing, China; Massachusetts Institute of Technology (MIT), Cambridge, MA, USA; The Ohio State University, Columbus, OH, USA; and École Normale Supérieure (ENS), Paris, France. His research has been supported by the Science and Technology Development Fund of Macau (FDCT), the National Natual Science Foundation of China (NSFC), and several companies including Huawei Technology, Horizon Robotics, and Amy Robotics. His research interests include sensing and perception for autonomous driving, mobile robotics, SLAM, and multiview geometry in computer vision.

Dr. Kong serves as an Associate Editor for the *International Journal of Computer Vision* (IJCV) and the *International Journal of AI and Autonomous Systems* (AIAS).