remote sensing

MDPI

Article

Multiscale and Multitemporal Road Detection from High Resolution SAR Images Using Attention Mechanism

Xiaochen Wei [1,2,3], Xikai Fu [1,2,*], Ye Yun [1,2] and Xiaolei Lv [1,2,3]

[1] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China; weixiaochen19@mails.ucas.ac.cn (X.W.); yunye@aircas.ac.cn (Y.Y.); lvxl@aircas.ac.cn (X.L.)
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: xkfu@mail.ie.ac.cn

Abstract: Road detection from images has emerged as an important way to obtain road information, thereby gaining much attention in recent years. However, most existing methods only focus on extracting road information from single temporal intensity images, which may cause a decrease in image resolution due to the use of spatial filter methods to avoid coherent speckle noises. Some newly developed methods take into account the multi-temporal information in the preprocessing stage to filter the coherent speckle noise in the SAR imagery. They ignore the temporal characteristic of road objects such as the temporal consistency for the road objects in the multitemporal SAR images that cover the same area and are taken at adjacent times, causing the limitation in detection performance. In this paper, we propose a multiscale and multitemporal network (MSMTHRNet) for road detection from SAR imagery, which contains the temporal consistency enhancement module (TCEM) and multiscale fusion module (MSFM) that are based on attention mechanism. In particular, we propose the TCEM to make full use of multitemporal information, which contains temporal attention submodule that applies attention mechanism to capture temporal contextual information. We enforce temporal consistency constraint by the TCEM to obtain the enhanced feature representations of SAR imagery that help to distinguish the real roads. Since the width of roads are various, incorporating multiscale features is a promising way to improve the results of road detection. We propose the MSFM that applies learned weights to combine predictions of different scale features. Since there is no public dataset, we build a multitemporal road detection dataset to evaluate our methods. State-of-the-art semantic segmentation network HRNetV2 is used as a baseline method to compare with MSHRNet that only has MSFM and the MSMTHRNet. The MSHRNet(TAF) whose input is the SAR image after the temporal filter is adopted to compare with our proposed MSMTHRNet. On our test dataset, MSHRNet and MSMTHRNet improve over the HRNetV2 by 2.1% and 14.19%, respectively, in the IoU metric and by 3.25% and 17.08%, respectively, in the APLS metric. MSMTHRNet improves over the MSMTHRNet(TAF) by 8.23% and 8.81% in the IoU metric and APLS metric, respectively.

Keywords: road detection; attention mechanism; deep learning; multi-temporal; multi-scale; SAR imagery

Citation: Wei, X.; Fu, X.; Yun, Y.; Lv, X. Multiscale and Multitemporal Road Detection from High Resolution SAR Images Based Attention Mechanism. *Remote Sens.* **2021**, *13*, 3149. https://doi.org/10.3390/rs13163149

Academic Editor: Dusan Gleich

Received: 12 July 2021
Accepted: 4 August 2021
Published: 9 August 2021

1. Introduction

The road information is essential in various practical applications, such as urban planning, traffic measurement, auxiliary navigation, GIS database update, and emergency response [1,2]. With the help of computer technology, automatically extracting roads from remote sensing images becomes economical and effective. Especially, synthetic aperture radar (SAR) has received a lot of attention in the road extraction area recently due to the wide coverage and day-and-night and all-weather observation capability [3–9].

In the high-resolution SAR imagery, the roads may be precisely modeled as dark elongated areas surrounded by bright edges, which are due to double-bounce reflections by surrounding buildings or uniform backscattering by the vegetation [10]. However, in practical, there are various interferences that includes speckle noise, layover and shadow make the road detection be a non-trivial work.

In the past few decades, researchers have proposed various methods to automatically extract roads from SAR images. Most of methods consist of two main steps: road detection and road network reconstruction. Tupin et al. in [11] modified the MRF model by adding the clique potentials to adapt the road network extarction to refine [6]. Negri et al. in [10] proposed a multiscale feature fusion detector that extracted multiscale radiance, direction, edge feature, and mended the MRF model in [11], considering two kinds of nodes (T-shaped nodes and L-shaped nodes) in the road network and altering the selection of the main parameters of the MRF optimization chain. He et al. in [12] firstly established a multiscale pyramid on the input image and subdivided image for each level into a series of binary squares forming a quadtree, and then employed the multi-scale linear feature detector and beamlet to detect roads and adopted concept of regional growth to construct a road network, and proposed a rapid parameter selection procedure for adaptive adjustment of growth parameters. Lu et al. in [3] proposed a weighted ratio line detector (W-RLD) to extract the faeture of the road and used an automated road seed extraction method that combines ratio and direction information to improve the quality of road detection. The above methods separately executed two steps, which easily incur error accumulation. The work [5] achieved better performance using a conditional random field model to simultaneously perform the two steps.

With the a rise of deep learning, many works [13–15] for road extraction using deep convolutional neural networks (DCNN) have emerged, which greatly promote the improvement of road extraction performance. The fully convolutional neural network (FCN) [16] that stacks several convolutional and associative layers to gradually expand the receptive field of the network is easier to obtain road information [17,18]. The U-Net architecture [19] that connects feature maps with different resolutions was employed to extarct road information [20–23]. D-LinkNet [24] with encoder-decoder architecture captures the rough position of the road extraction through the high level features with large receptive field and refines the edge through the high resolution feature maps that retain the spatial structure details. Xu et al. in [25] proposed a road extraction neural network based on spatial attention, which can capture the context information between roads and buildings to extract roads more accurately. RoadNet [26] predicted road surface, edge and centerline at the same time under multitask learning scheme. So far, deep learning is mainly applied to road extraction from optical remote sensing images. Less attention are paid on extracting road information from SAR imagery via deep learning. This may due to the fact that the unique characteristics of SAR images make labeling time-consuming and labor-intensive. Although there have been fewer minimal works that apply deep learning for road extraction from SAR imagery to date, it is no doubt that deep learning has colossal application potential for extraction of road information from SAR image. Henry et al. in [7] applied Deep Fully Convolutional Neural Networks (DFCNs) to segment roads from SAR imagery, which achieved good performance. Wei et al. in [8] proposed a multitask learning framework that learn road detection task and road centerline extraction based ordinal regression task simultaneously and designed a new loss named road-topology loss to improve the connectivity and complement of road extraction results.

Along with the development of modern acquisition technology that produces the high-resolution SAR data stream with a short repetition period, there have been some works using multitemporal SAR images for road extraction. Chanussot et al. in [27] designed a directional prefilter that uses majority voting and morphological filters to adaptively explore the possible directions of linear structures on temporal averaged SAR images. For robust road extraction, ref. [9] estimated coherence without loss of image resolution by homogeneous pixel selection and robust estimators. Then, ref. [9] combined coherence

and temporal averaged intensity to detect roads. In [9,27], multitemporal information is only used in the preprocessing stage to despeckle, which loses a lot of useful temporal information. The roads in the remote sensing imagery change slowly, and the roads neither suddenly appear nor disappear suddenly [28]. In a relatively short period of time, there are temporal correlation between the roads in the SAR images acquired at different times. Futhermore, the speckle noise in SAR images is random multiplicative noise. As a result, SAR images acquired at different times can complement each other to reduce the interference of coherent speckle noise [29]. These inspire us to exploit multitemporal information for better road detection performance. In this paper, we propose a temporal consistency enhancement module that uses temporal attention mechanism to capture long-range temporal context information, by which the representation with temporal consistency of SAR imagery is obtain.

In real world, the width of roads varies greatly. There exists a trade off in road detection task that the road with smaller width of predictions are best handled at lower inference resolution and predicting other roads with larger width are better handled at higher inference resolution. Fine detail, such as the boundary is often better predicted with high resolution feature maps. And at the same time, predictions of large roads, which requires more global context, is often done better with large receptive field. As a result, multiscale feature maps are useful for road detection task. A lot of works [23,30–32] have emerged in recent years to address this trade off. Lu et al. in [30] proposed a multi-scale enhanced road detection framework (DenseUNet), which employed atrous spatial pyramid pooling (ASPP) to effectively capture multi-scale features for road detection. In paper [31], multiscale convolution that can provide higher accuracy is applied to obtain hierarchical features with different dimensions. Lu et al. in [23] employed multiscale feature integration to combine features from different scales to improve the robustness of feature extraction. In paper [32], the proposed network can produce intermediate outputs of different scales and used multiscale losses to guide the network during training. The pervious works [23,30,31] just leverage features from the final layer of network to detect roads. The paper [32] penalize the prediction of each scale for each object equally, which ignores the importance of prediction results at different scales for each objects is different. In this paper, we leverage multiscale features instead of just the features from the final layers of the network to obtain multiscale predictions, and propose a multiscale fusion module to combine multiscale predictions by weight for each scale learned by scale attention module.

Our main contributions are as follows:

1. For multitemporal road detection from SAR imagery, we build our dataset with TerraSAR-X images, which cover same areas and were obtained at different times. Our experiments are carried out based on this dataset. Our experimental results show that our proposed framework can achieve a better road detection performance;
2. In this paper, an temporal consistency enhancement module is proposed to obtain the representation with temporal consistency under temporal attention mechanism that is used to capture long range temporal context information;
3. We propose an efficient multi-scale fusion module that merges predictions of feature maps with different receptive fields by learning weights for different scales, which helps predict roads with various width.

2. Materials and Methods

In this section, we present the details of the proposed multiscale and multitemporal network (MSMTHRNet) for road detection from SAR images. We first present a general framework of our multitemporal multiscale network. Then, the temporal consistency enhancement module which can enforce temporal consistency between features of different temporal SAR images by capture of temporal contextual information is proposed. Finally, we present the multiscale fusion module that can combine predictions of different resolution feature maps.

2.1. Overview

We denote by $(I_0, ..., I_c, ..., I_{T-1})$ a sequence of SAR images that cover the same area and shot at different times. Only SAR image I_c has a corresponding road groundtruth, which is referred to as the main image. Except for SAR image I_c, the remaining images are referred to as auxiliary images. Our proposed multitemporal multiscale road detection network use the multitemporal SAR images as input and aims to detect roads in SAR image I_c.

As shown in Figure 1, the multitemporal SAR images are firstly fed into a shared backbone to extract features, which directly follows the work [33]. Ref. [33] proposed a network architecture HRNet, which has achieved better performance in various applications including semantic segmentation, pose estimation, etc. Different with the frameworks that frist encode the input image as a low-resolution representation through a subnetwork that is formed by connecting high-to-low resolution convolutions in series (e.g., ResNet, VGGNet) and then recover the high-resolution representation from the encoded low-resolution representation, the HRNet maintains high-resolution representations through the whole process. As shown in Figure 2, the HRNet has two key characteristics. One is that the HRNet connects the high-to-low resolution convolution streams in parallel, and the other is that HRNet can repeatedly exchange the information across resolution. The resulting representation of the HRNet is semantically richer and spatially more precise, which is essential for road detection problems. As a result, we choose the HRNet as the backbone to extract features.

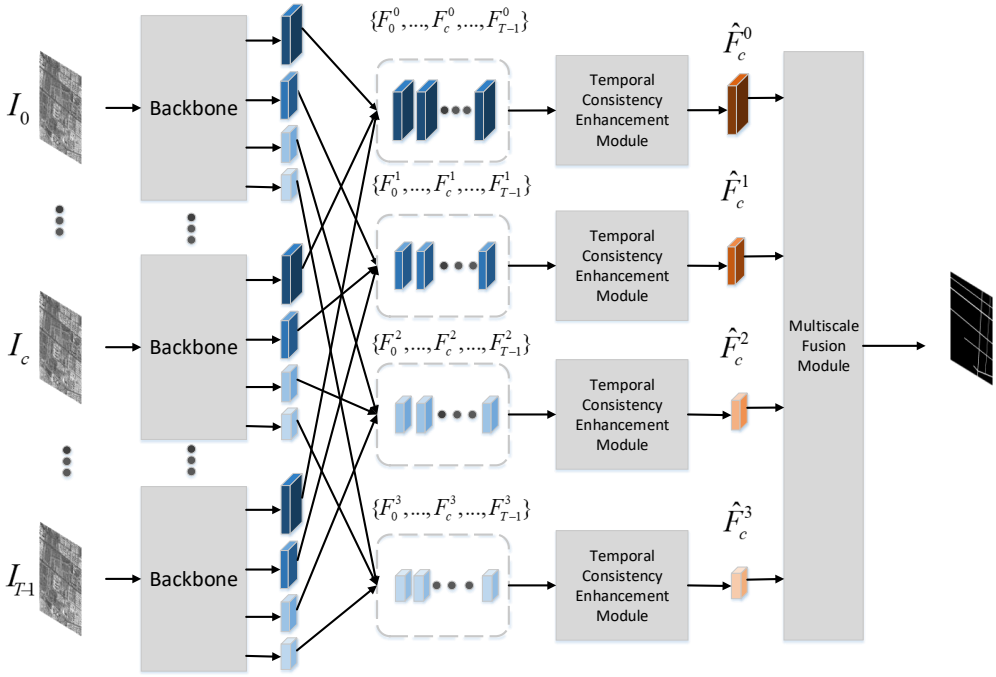

Figure 1. The overview of Multiscale and Multitemporal Road Detection Framework.

The HRNet produces multitemporal multiscale feature maps $\{F_0^0, ..., F_0^s, ..., F_0^3\}, ..., \{F_c^0, ..., F_c^s, ..., F_c^3\}, ..., \{F_{T-1}^0, ..., F_{T-1}^s, ..., F_{T-1}^3\}$, where F_t^s is the feature map of SAR image I_t for s scale. Multitemporal feature maps with the same resolution $\{F_0^s, ..., F_c^s, ..., F_{T-1}^s\}$ are input to the temporal consistency enhancement module, which can capture long-range temporal context information and enhance the feature representation of the main image by enforcing temporal consistency constraints. After the temporal consistency enhancement module, the multiscale feature maps $\{\hat{F}_c^0, ..., \hat{F}_c^s, ..., \hat{F}_c^3\}$ with temporal consistency of main image I_c are obtained. Finally, the multiscale fusion module combines predictions of

multiscale feature maps $\{\hat{F}_c^0, ..., \hat{F}_c^s, ..., \hat{F}_c^3\}$ by learning the weights of feature maps for different scales.

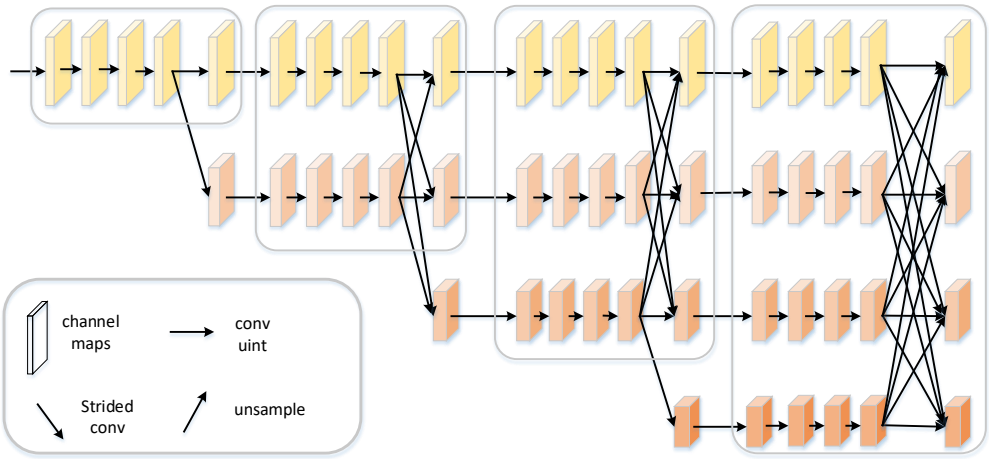

Figure 2. The architecture of High-Resolution Network. There are four stages. The first stage consists o high-resolution convolutions. The second, third and fourth stages repeat two-resolution, three-resolution and four-resolution blocks, respectively. HRNet can connect high-to-low convolution streams in parallel and maintain high-resolution features through the whole process [33].

There are abundant temporal information in multitemporal SAR images, which are essential for road detection from SAR images. There are two reasons that motivate us to use multi-temporal information to detect roads from SAR images. One is that the roads in the remote sensing imagery are slowly changing objects, which neither suddenly appear nor disappear. Especially, in a very short period of time, the roads in the remote sensing imagery could be seen as remaining unchanged. There is temporal consistency between the roads in remote sensing images that are taken at adjacent temporal and cover the same area. The other is that coherent speckle noise in SAR images is random multiplicative noise and multitemporal SAR images covering the same area can complement each other. Huang et al. [34] proposed a criss-cross attention module to capture full-image spatial contextual information. Huang et al. [34] used two criss-cross attention module that only need lightweight computation and memory to replace non-local attention module. Given local feature representations H, the criss-cross attention module firstly applies two 1×1 convolutions to generate three dimensional feature maps Q , K, and V. And Then Q and K are be fed into affinity operation to calculate spatial attention map. Finally, dot multiplication result of the spatial attention map A and V add H to generate new feature maps H'. As a result, inspired by criss-cross attention module [34] that is used to model the full-image spatial dependencies, we propose a temporal consistency enhancement module to impose temporal consistency constraints by capturing long-range temporal contextual information.

Figure 3 illustrates our proposed temporal consistency enhancement module that uses multitemporal feature maps for s scale $\{F_0^s, ..., F_c^s, ..., F_{T-1}^s\}$ as input and outputs temporal enhanced feature representation of main SAR image I_c, i.e., $\hat{F}_c^s$. Each $F_t^s (t = 0, ..., c, ...T-1)$ firstly fed into two encoder layers. The first encoder layer is used to encode different representations $\{\tilde{F}_0^s, ..., \tilde{F}_c^s, ..., \tilde{F}_{T-1}^s\}$ of differnet temporal SAR images, which is composed of one 1×1 convolution and one 3×3 convolution. Directly using original output of backbone is computationally expensive because of the high-dimensional channel. As a result, the second encoder layer is adopted to channel reduction by encoding key features $\{K_0^s, ..., K_c^s, ..., K_{T-1}^s\}$ of different temporal SAR images. We separately enhance the temporal consistency of each temporal feature map.

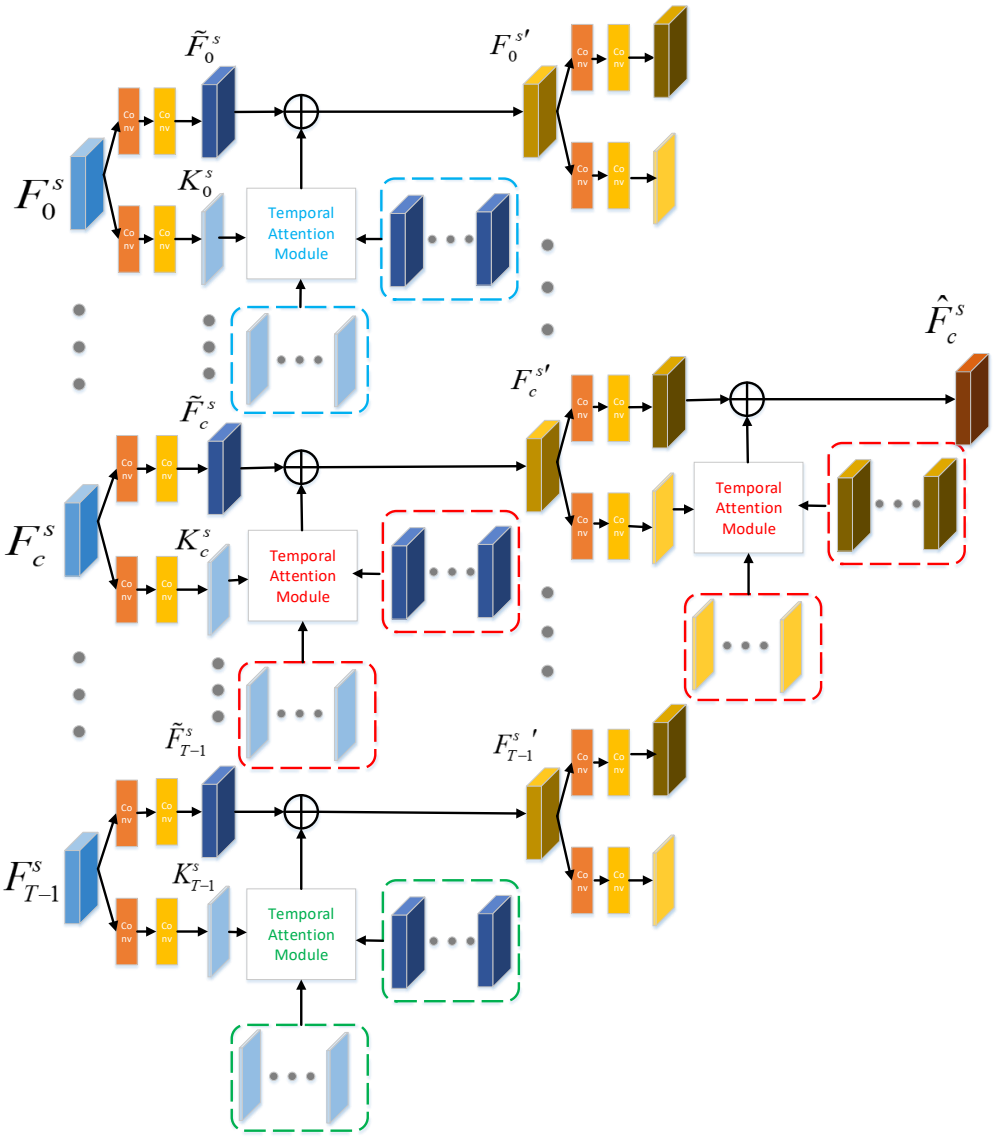

Figure 3. Temporal Consistency Enhancement Module.

2.2. Temporal Consistency Enhancement Module

For feature map $\tilde{F}_t^s$ of SAR image I_t, we firstly adopt our proposed temporal attention module to capture long-range temporal contextual information. Except for the feature map $\tilde{F}_t^s$, the feature maps of other temporal SAR images and the key features of all temporal are input into temporal attention module.

As depicted in Figure 4, we concatenate the feature maps and key features of other temporal SAR images along the temporal dimension generating a 4-dimension matrix, and permute them to $M_F^t \in R^{H_s \times W_s \times C_F^s \times (T-1)}$ and $M_K^t \in R^{H_s \times W_s \times (T-1) \times C_K^s}$, respectively. We permute and reshape the key feature of SAR image I_t to $Q_K^t \in R^{H_s \times W_s \times C_K^s \times 1}$. Next, we multiply M_K^t and Q_K^t, and then apply a softmax layer to calculate the temporal attention map $A_t \in R^{H_s \times W_s \times (T-1) \times 1}$. Especially, $A_t^{t'} \in A_t$ is the degree of correlation between key feature K_t and $K_{t'}$ and is given as:

$$A_t^{t'} = \frac{\exp(K_t \cdot K_{t'})}{\sum_{t'' \neq t} \exp(K_t \cdot K_{t''})}, \tag{1}$$

where $A_t^{t'}(i,j)$ measures the impact of the (i,j) position in the key feature K_t on the same position in the key feature $K_{t'}$. It should be noted that the larger impact from the I_t to the $I_{t'}$, the greater relation between them. After obtaining the correlation map A_t, we multiple A_t and M_F^t to combine temporal relation with multitemporal features. Finally, we reshape and permute $\tilde{M}_F^t$ to obtain temporal attention feature $\overline{\tilde{F}}_t^s$.

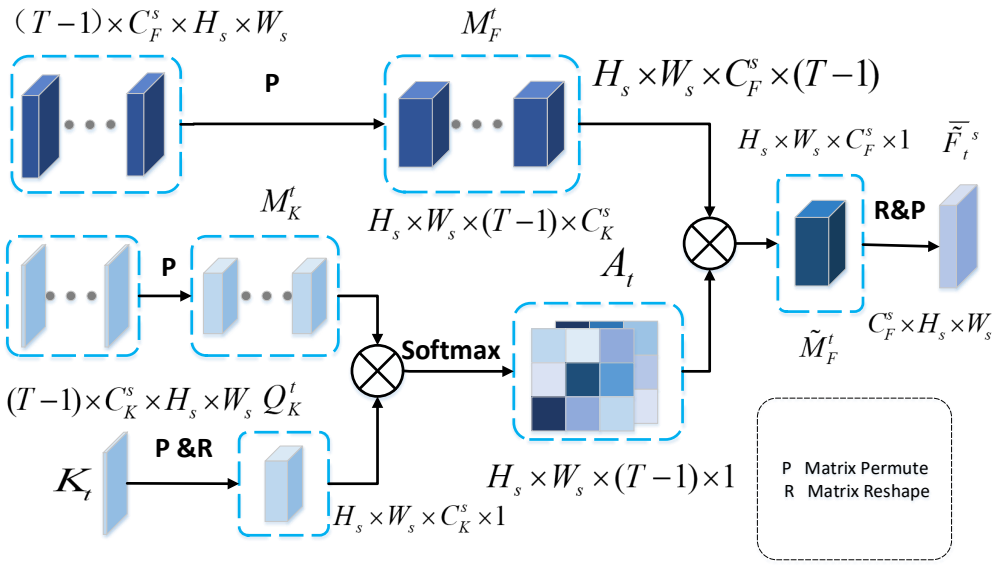

Figure 4. Temporal Attention Module.

As shown in Figure 3, after obtaining the long-range temporal context information via temporal attention module, we add the temporal attention feature $\overline{\tilde{F}}_t^s$ and feature $\tilde{F}_t^s$ to enhance the representation of image I_t. The aggregation feature $F_t^{s'}$ is given by:

$$F_t^{s'} = \overline{\tilde{F}}_t^s + \tilde{F}_t^s. \tag{2}$$

After the first round temporal consistency enhancement, we obtain the multitemporal feature representations with temporal consistency for s scale, i.e., $\{F_0^{s'}, ..., F_c^{s'}, ..., F_{T-1}^{s'}\}$. In the second round, we use multitemporal feature maps $\{F_0^{s'}, ..., F_c^{s'}, ..., F_{T-1}^{s'}\} \setminus F_c^{s'}$ ($\setminus$ is minus operation of set) to enhance the representation of main image. As shown in Figure 3, the output of temporal consistency enhancement module is the aggregation feature of main image I_c for s scale, i.e., $\hat{F}_c^s$.

2.3. Multiscale Fusion Module

In practice, low-resolution feature map with large receptive field is easier to predict larger objects in a scene. However, due to multiple downsampling operations, it is difficult to detect small objects. High-resolution feature map with small receptive field is just the opposite. Due to the less down-sampling operations, high-resolution feature map retains more details, which resolves fine detail better. In the real world, the width of roads varies greatly. Low-resolution feature map with larger receptive field have difficulty to predict narrow roads and ensure accurate road edges. High-resolution feature map with smaller receptive field can not distinguish between wide roads and rivers. In order to make full use of the advantages of different resolution feature map, we propose a multiscale fusion module based on the scale attention mechanism that is similar to the paper [35]. As shown in the Figure 5, feature map for the s scale of main SAR image $F_c^s (s = 0, 1, 2, 3)$ is fed into the corresponding segmentation head to obtain the road detection prediction map y^s for the scale s. The segmentation head for each scale is the same, which consists of a 3×3 convolution layer, BatchNorm layer, ReLU activation layer, a 1×1 convolution layer and a

Softmax activation layer. When $s = 1, 2, 3$, we need to upsample $\hat{Y}^s$ to guarantee that the size of road prediction map for each scale is identical.

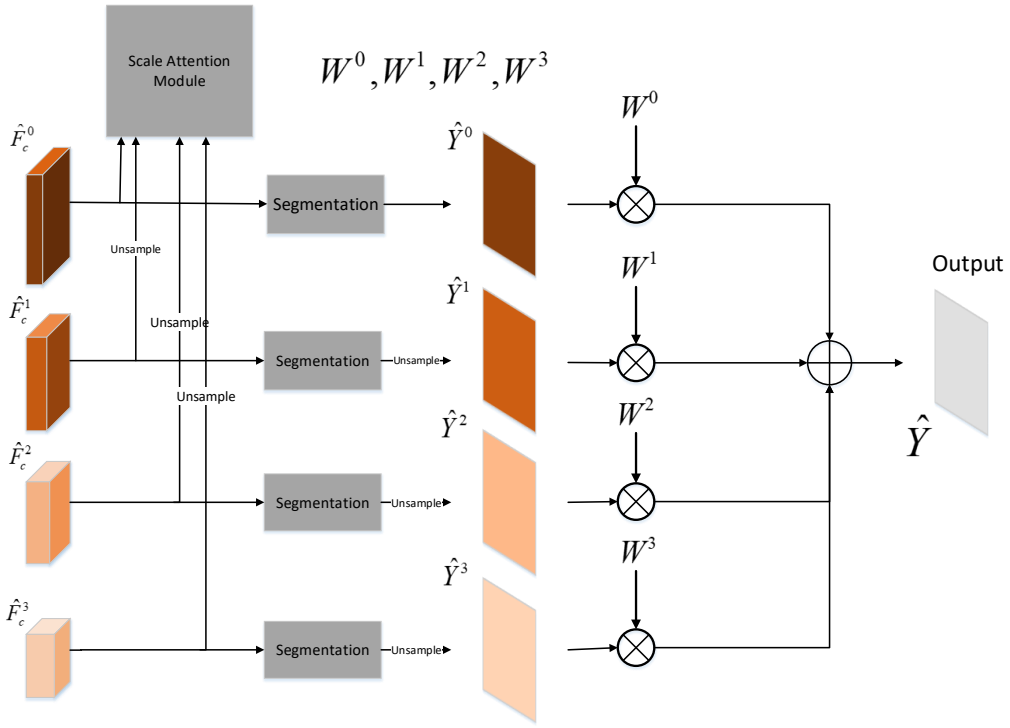

Figure 5. Multiscale Fusion Module.

As shown in Figure 6, the scale attention module feds with the multiscale feature maps $\{\hat{F}_c^0, \hat{F}_c^1, \hat{F}_c^2, \hat{F}_c^3\}$. We upsample $\hat{F}_c^1$, $\hat{F}_c^2$ and $\hat{F}_c^3$, so that these feature maps have the same width and height with the feature map $\hat{F}_c^0$. As shown in Figure 6, the multiscale feature maps are input to the scale attention module to learn the weights corresponding to each scale $\{W^0, W^1, W^2, W^3\}$. The scale attention module is composed of $(3 \times 3conv) \rightarrow (1 \times 1conv) \rightarrow (Softmax)$. The weight $W^s (s = 0, 1, 2, 3)$ is computed by the Equation (3).

$$W^s = \frac{exp(G^s)}{\sum_{s'=0}^{4} exp(G^{s'})}, \tag{3}$$

where G^s is last layer output before SoftMax produced by the scale attention module for scale s.

The final prediction $\hat{Y}$ is the weighted sum of $\hat{Y}^s$ for all scales, i.e.,

$$\hat{Y} = \sum_{s=0}^{4} W^s \circ \hat{Y}^s, \tag{4}$$

where $\circ$ is Hadamard multiply.

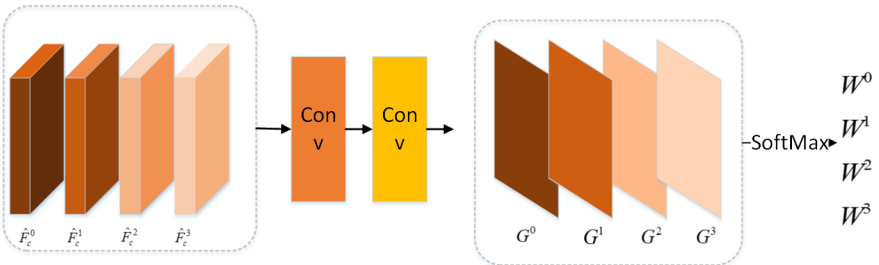

Figure 6. Scale Attention Module.

3. Experiment

3.1. Dataset

In this subsection, we aim to introduce the dataset used in this paper. There is no public data set suitable for our research. We create our dataset using high-resolution TerraSAR-X images obtained by the stripe mode. As shown in Table 1, we select two study areas, which contain urban, suburban, and rural regions. There are seven SAR images that are obtained at different times for each study areas. We split raw SAR images to patches, the size of which are 1024×1024. In other SAR road detection datasets, the size of patch is usually 256×256 [7]. The reason why we split the raw SAR images to 1024×1024 is to be able to capture a longer range of spatial context information, which is very important for road detection task. We remove the samples, which do not have road regions or have few road regions. In this way, we obtain 495 samples in our dataset. We randomly select 397 samples as our train dataset and 98 samples as our test dataset. Our train set contains multitemporal SAR patches and the groundtruth of main SAR patches. For the first study area, the main SAR images are obtained at 9 May 2013. For the second study area, the main SAR images are obtained at 4 March 2013.

Table 1. Metadata of the TerraSAR-X Images Used in Our Data Set.

Area1	**Beijing, China**
Size	21,800 × 15,500 px
Range Sample Distance	0.909627 m/px
Azimuth Sample Distance	1.848561 m/px
Spatial Resolution	3 m/px
Center Coordinate	[39.8798466, 116.4503446]
Polarization	HH
Date	19 January 2013, 21 February 2013, 26 March 2013, 9 May 2013, 31 May 2013, 3 July 2013, 5 August 2013
Area2	**Beijing, China**
Size	27,600 × 18,700 px
Range Sample Distance	0.908790 m/px
Azimuth Sample Distance	1.888833 m/px
Spatial Resolution	3 m/px
Center Coordinate	[39.957164, 116.6996268]
Polarization	HH
Date	11 May 2012, 20 September 2012, 23 October 2012 4 March 2013, 22 June 2013, 14 July 2013, 16 August 2013

3.2. Metric

3.2.1. Pixel-Based Metrics

To evaluate the performance of our method for road detection from a pixel perspective, we adapt the metrics [36] including precision (P), recall (R), Intersection over Union (IoU), and pixel-based F1-score denoted as $F1_P$. Pixel-based precision (P) measures the ratio of the number of the pixels which are labeled as road pixels in the ground truth and are predicted as road pixels to the number of pixels that are inferred as road pixels. Pixel-based recall (R) calculates the ratio of the number of the pixels which are labeled as road pixels in the ground truth and are predicted as road pixels to the number of pixels that are labeled as road pixels. $F1_P$ is used to balance precision and recall, which is a harmonic average between precision and recall. Intersection over union (IoU) is the ratio of the intersection of prediction and groundtruth to the union of prediction and groundtruth, which can trade-off between recall and precision. Specifically, the four metrics are defined as:

$$P = \frac{TP}{TP + FP}, \tag{5}$$

$$R = \frac{TP}{TP + FN}, \tag{6}$$

$$IoU = \frac{TP}{TP + FN + FP}, \tag{7}$$

$$F1_P = \frac{2 \times P \times R}{P + R}, \tag{8}$$

where TP is true positive, FP is false positive and FN is false negative. Since there is a deviation between the manually labelled roads and the real roads, we relax metrics using the buffer method given in [37]. Specifically, if the regions in the prediction result are within the two pixels range, they are regarded as matching regions.

3.2.2. Topology-Based Metric

Pixel-based metrics are universal metrics for segmentation tasks, which are not enough for evaluating the performance of roads detection results. Pixel-based metrics evaluate the prediction results of each pixel equally. The loss of a few pixels has little effect on road segmentation results, but it does have a significant effect on road connectivity. As a result, we adopt three topology-based metric: the average path length similarity ($APLS$) [38] and Biagioni F1-score ($F1_{Bi}$) [39] to measure the estimated topology and road connectivity. These two metrics will be described below.

The first metric $APLS$ captures the deviation of the shortest path distance between all of the node pairs in the graph. We obtain the groundtruth graph G and predicted road network graph $\hat{G}$ from Y and $\hat{Y}$, respectively. $S_{P \to T}$ measures the sum of difference of shortest path for each node pair in groundtruth graph $G = (V, E)$ and estimated graph $\hat{G} = (\hat{V}, \hat{E})$. To penalize the positives, symmetric term $S_{T \to P}$ is added to $APLS$ metric which considers predicted graph as groundtruth and true graph as prediction.

$$S_{P \to T} = 1 - \frac{1}{|V|} \sum \min\left(1, \frac{|L(a,b) - L(\hat{a},\hat{b})|}{L(a,b)}\right) \tag{9}$$

$$APLS = \frac{1}{N} \sum_{(y,\hat{y})} \left(\frac{1}{\frac{1}{S_{P \to T}(G,G)} + \frac{1}{S_{T \to P}(G,G)}}\right) \tag{10}$$

where $a, b \in V$, $\hat{a}, \hat{b} \in \hat{V}$, $|V|$ is the number of nodes in groundtruth graph, and N is the number of images in a minibatch. $L(a,b)$ and $L(\hat{a},\hat{b})$ are path length of $a \to b$ and $\hat{a} \to \hat{b}$, respectively.

The second metric $F1_{Bi}$ compares the sets of accessible locations by moving a predetermined distance away from the corresponding points in the two graphs. To this end, a starting position is randomly selected in the ground truth network. The closest point in the prediction network is identified. Then, the local subgraphs are extracted by breadth-first exploration of the graphs far away from the starting position. The calculation of the $F1_{Bi}$ is based on spatial coincidence "Control points" are inserted into the subgraph at regular intervals. The control point in ground truth is called hole. The control point in predicted network is called marble. If a marble is close enough to a hole in predicted network, the marble is considered to be matched marble. If a hole is close enough to a marble in ground truth network, the hole is considered to be matched hole. Control points that do not match in the predicted and annotated subgraphs are considered spurious marbles and empty holes, respectively. The sampling and matching of the local subgraph are repeated many times, and the spurious and missing are calculated based on the total number of matched and unmatched control points. According to spurious and empty, the $F1_{Bi}$ is calculated as:

$$F1_{Bi} = 2 \times \frac{(1 - spurious) + (1 - missing)}{(1 - spurious) \times (1 - missing)} \tag{11}$$

where:

$$spurious = \frac{spuriousmarbles}{spuriousmarbles + matchedmarbles}, \tag{12}$$

$$missing = \frac{emptyholes}{emptyholes + matchedholes} \tag{13}$$

3.3. Implementation Details

We adopt the Pytorch framework to implement networks trained on a single NVIDIA Tesla V100 with 16G memory using batch size of one. We train the networks with AdamW optimizer with the initial learning rate of 1.0×10^{-3}. To improve the robustness of network, we apply data augmentation for images in our train dataset including random horizontal flip, random vertical flip, and random rotation 90 degree. There are 397 SAR images in the train dataset. After data augmentation, there are 3176 SAR images in train dataset. Roads are the objects whose areas must be larger than a certain number of pixels that can be set according to the spatial resolution. As a result, in the postprocessing stage, we discards the small-area regions whose areas smaller than eighty pixels to improve the performance of road detection results.

3.4. Results

In this subsection, we will firstly introduce the networks to verify the effectiveness of the MSFM and TCEM. And then we will show the performance of baseline method and our proposed methods.

3.4.1. Baseline and Variants of the Proposed Method

We choose the HRNetV2 [33] as benchmark to study the performance of multiscale fusion module (MSFM). As shown in Figure 7, we modify the HRNet with multiscale fusion module to obtain the network MSHRNet. Next, we adopt the image after temporal average filter as input of MSMTHRNet to study the performance of temporal consistency enhancement module (TCEM). As shown in Figure 7, we denoted this method as MSMTHRNet(TAF).

3.4.2. Comparative Evaluation

To compare the performance of road detection, all the methods are evaluated based on the test samples in the test set for road detection.

We present quantitative comparisons in Table 2. According to Table 2, MSHRNet outperforms the HRNetV2, this is due to the modified MSFM. By adding TCEM, our proposed MSMTHRNet achieves better performance than MSMTHRNet(TAF). The reason is that the TCEM can capture long range temporal contextual information to obtain enhanced representation. The differences are greater when using the metric specifically designed to gauge the quality of road network reconstruction. For pixel-based metric IoU, MSHRNet and MSMTHRNet improve over the HRNetV2 by up to 2.1% and 14.19%, and MSMTHRNet improve over the MSHRNet (TAF) by up to 8.23%. Form the Table 2, we can discover that the MSMTHRNet improves the APLS by 17.08% and 8.81% over the HRNetV2 and MSHRNet(TAF) respectively, which reveals that our proposed method can improve road connectivity. For topology-based metric TOSM, our proposed method greatly improves the correctness of road topology and decrease the number of infeasible paths that indicates missing links.

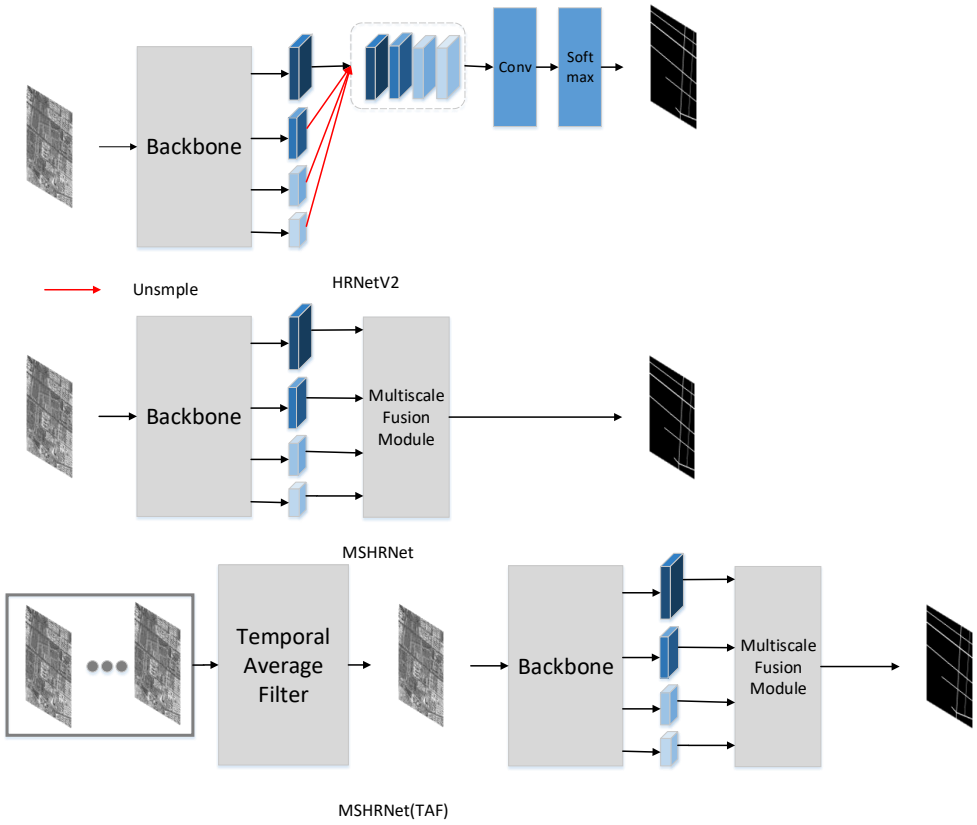

Figure 7. HRNetV2 is baseline method; MSHRNet is modify the baseline method with multiscale fusion module. MSMTHRNet(TAF) use the image after temporal average filter.

Table 2. Comparative quantitative Evaluation Among Different Methods for Road Detection on our dataset. It should be noted that the results are the average performance of all images in the test set. A higher value indicates a better performance. With the best results marked in bold.

	Pixel-Based Metrics			
	P	R	IoU	F1
HRNetV2	**0.9383**	0.6367	0.6125	0.7488
MSHRNet	0.9113	0.6751	0.6335	0.7662
MSHRNet(TAF)	0.9352	0.7038	0.6721	0.7972
MSMTHRNet	0.9252	**0.8011**	**0.7544**	**0.8549**
	Topology-Based Metrics			
	APLS		$F1_{Bi}$	
HRNetV2	0.3730		0.5013	
MSHRNet	0.4055		0.5301	
MSHRNet(TAF)	0.4557		0.5624	
MSMTHRNet	**0.5438**		**0.6529**	

For qualitative comparison, we present the results produced by all methods based on example images depicted in Figures 8 and 9. There are both six columns of subfigures in Figures 8 and 9. The first, third and fifth columns illustrate the results of three sampled images, and the second, fourth, and sixth columns illustrate the close-ups of the corresponding regions of red rectangles in the first, third, and fifth columns, respectively. It can be seen from Figure 8 that our proposed methods detect more true road structures and eliminates false negative in the foreground without predicting too much spurious road regions. From Figure 9, we can discover that the road detection result of MSMTHRNet

have better road connectivity than the prediction of MSHRNet(TAF) by applying temporal consistency enhancement module.

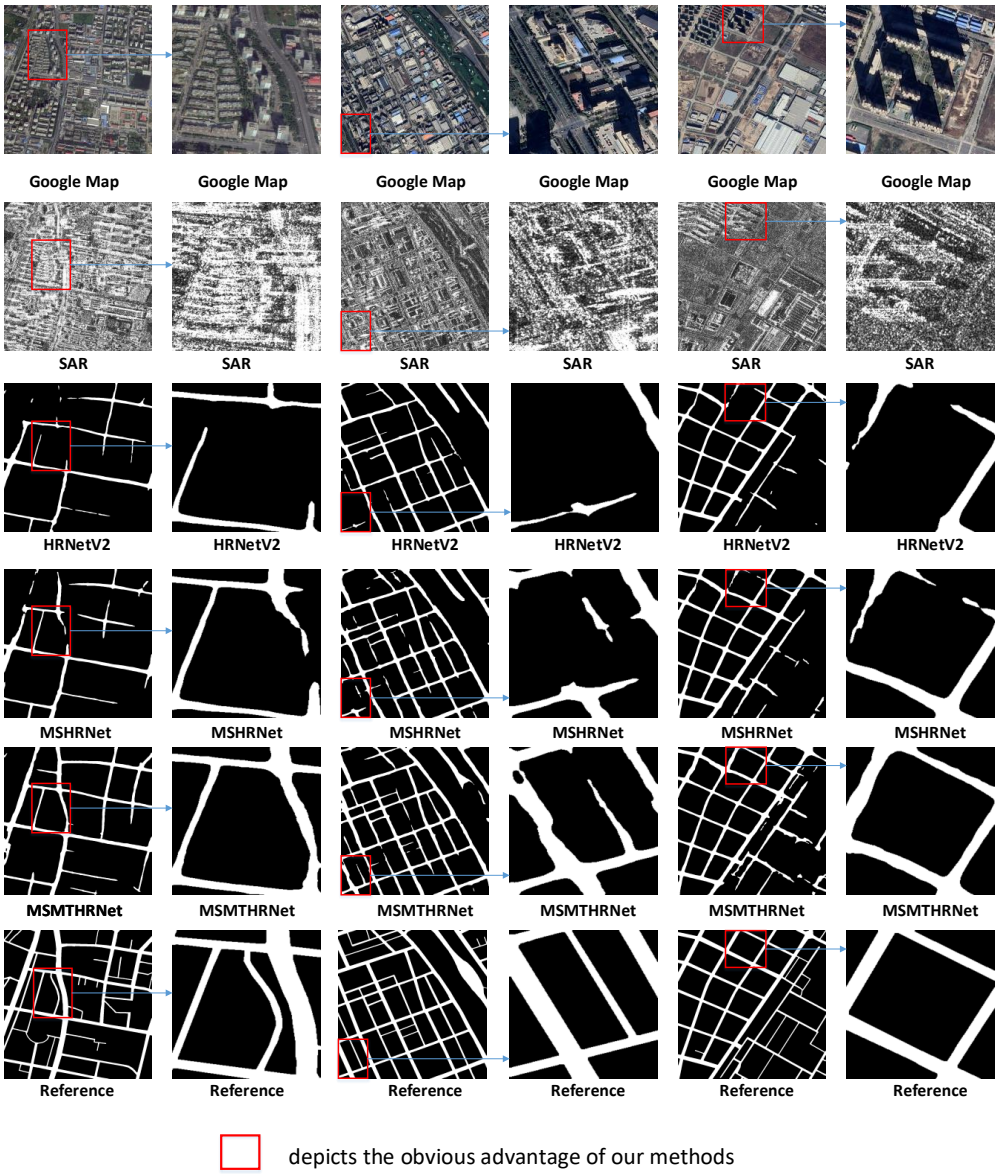

Figure 8. Qualitative comparison of road detection results produced by different methods based on three images from our test set.

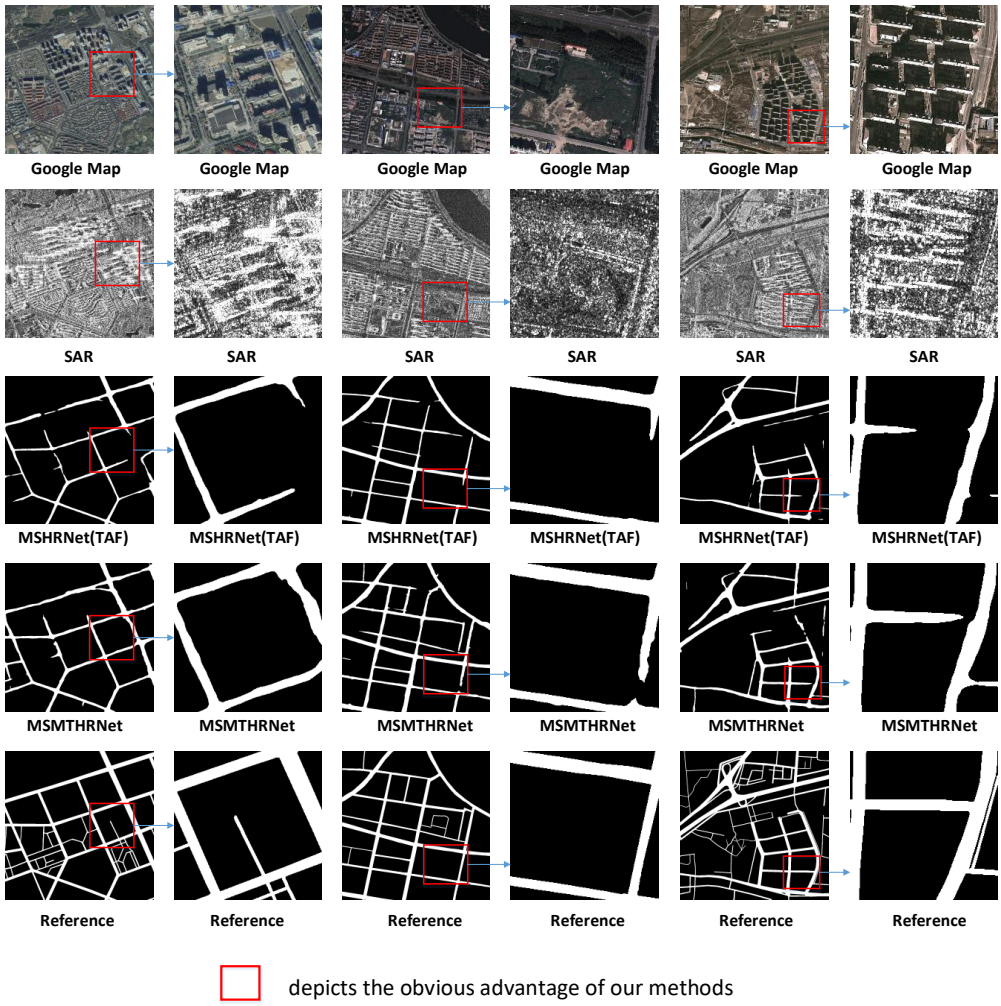

Figure 9. Qualitative comparison of road detection results produced by different methods based on three images from our test set.

4. Discussion

In this section, we will discuss how the number of input SAR images affects the performance of our proposed method. We evaluate the MSHRNet whose input is a single SAR image, the MSMTHRNet ($T = 3$) with three input SAR images, the MSMTHRNet ($T = 5$) with five input SAR images and the MSMTHRNet ($T = 7$) with seven input SAR images on the test samples in the test set. The quantitative comparisons are summarized in Table 3 and Figure 10. And the qualitative comparisons are illustrated in Figure 11. Form Table 3 and Figure 10 , we can see that by increasing the number of input SAR images, the performance of road detection continuously improved. From the pixel-wise perspective, MSMTHRNet ($T = 3$), MSMTHRNet ($T = 5$) , MSMTHRNet ($T = 7$) are 6.61%, 8.6% and 12.99% higher than the MSHRNet, respectively, in the IoU metric. From the topology-wise perspective, MSMTHRNet ($T = 3$), MSMTHRNet ($T = 5$), and MSMTHRNet ($T = 7$) are 9.48%, 11.78%, and 14.54% higher than the MSHRNet, respectively in the APLS metric, which reveals that the connectivity of road is enhanced continuously by increasing the input SAR images. As shown in Figure 11, increasing input SAR images makes the missing road regions consistently decrease and road detection results have better road connectivity. The reason is that the more SAR images input, the better the temporal consistency of the enhanced feature representation obtained through the temporal consistency enhancement module, which can reduce the affects of occlusion, coherent spots, and shadows.

Table 3. Comparative quantitative Evaluation Among Different Methods for Road Detection on our dataset. It should be noted that the results are the average performance of all images in the test set. A higher value indicates a better performance. With the best results marked in bold.

	Pixel-Based Metrics			
	P	R	IoU	F1
MSHRNet	0.9113	0.6751	0.6335	0.7662
MSMTHRNet (T = 3)	**0.9350**	0.7336	0.6992	0.8176
MSMTHRNet (T = 5)	0.8954	0.7837	0.7196	0.8320
MSMTHRNet (T = 7)	0.9252	**0.8011**	**0.7544**	**0.8549**
	Topology-Based Metrics			
	APLS		$F1_{Bi}$	
MSHRNet	0.4055		0.5301	
MSMTHRNet (T = 3)	0.4920		0.6177	
MSMTHRNet (T = 5)	0.5218		0.6514	
MSMTHRNet (T = 7)	**0.5438**		**0.6529**	

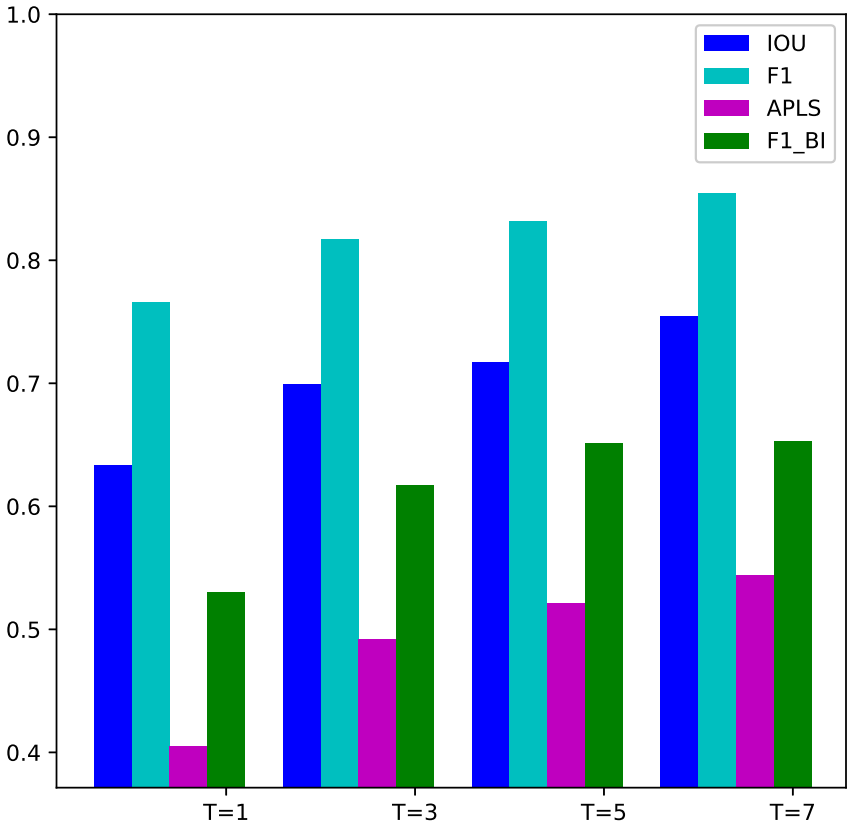

Figure 10. Various metrics IOU, F1-score, APLS, and $F1_{Bi}$ vs. the number of SAR input images.

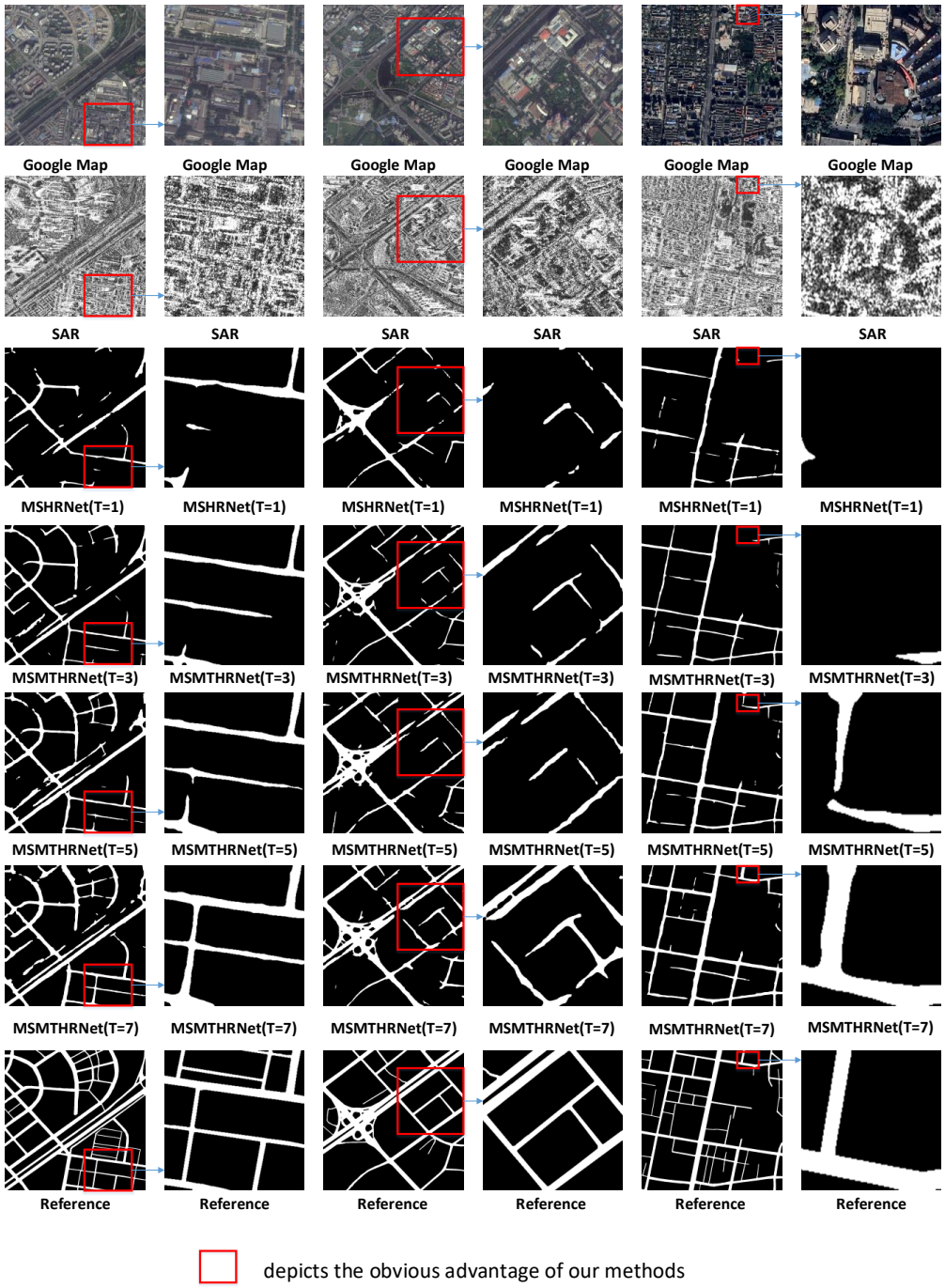

Figure 11. Qualitative comparison of road detection results produced by different methods based on three images from our test set.

5. Conclusions

This paper has proposed automatic road detection from SAR imagery exploiting multitemporal SAR images. The results reveal that our proposed method can achieve satisfactory results. For road segmentation, this paper adapts a state-of-the-art model HRNet as backbone to extract multiscale features. In order to take advantage of feature mapping at different scales, we have proposed a multiscale fusion module that combines the predictions of different scale features using the weights learned by the scale attention mechanism. The results show that our multiscale fusion module achieve better performance than baseline method. To make full use of multitemporal information, we have

proposed a temporal consistency enhancement module, which adopts temporal attention mechanism to capture long-range temporal context information. By using temporal consistency enhancement module, our network can make different SAR images obtained at different time complement each other. The experimental results demonstrate that the enhanced representation of main SAR imagery which is obtained by temporal consistency enhancement module can help improve the performance of our network in the pixel-based metrics and topology-based metrics. Our experimental results also display that the more multi-temporal images input, the better the road detection results.

Although, our proposed method have achieved better performance than previous methods, there are also a lot of missing roads specially the narrower ones. In the future, we will add a superresolution module to our model so that the narrower roads can be detected. In order to make our model achieve better generalization performance, we will continue to expand our dataset in the future by adding SAR images from other sensors.

Author Contributions: Conceptualization, X.W.; methodology, X.W.; software, X.W.; validation, X.W.; formal analysis, X.W.; investigation, X.W.; resources, X.F.; data curation, X.W. and X.F.; writing—original draft preparation, X.W.; writing—review and editing, X.W., X.L. and X.F.; visualization, X.W.; supervision, X.W. and X.F.; project administration, X.L. and X.F.; funding acquisition, X.L. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Natural Science Foundation of China, grant number 41801356, National Key Research and Development Program of China, grant number 2018YFC1505100, the China Academy of Railway Sciences Fund, grant number 2019YJ028.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Shi, W.; Miao, Z.; Debayle, J. An Integrated Method for Urban Main-Road Centerline Extraction From Optical Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3359–3372. [CrossRef]
2. Suchandt, S.; Runge, H.; Breit, H.; Steinbrecher, U.; Balss, U. Automatic Extraction of Traffic Flows Using TerraSAR-X Along-Track Interferometry. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 807–819. [CrossRef]
3. Lu, P.; Du, K.; Yu, W.; Wang, R.; Deng, Y.; Balz, T. A New Region Growing-Based Method for Road Network Extraction and Its Application on Different Resolution SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *7*, 4772–4783. [CrossRef]
4. Li, Y.; Zhang, R.; Wu, Y. Road network extraction in high-resolution SAR images based CNN features. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 1664–1667.
5. Xu, R.; He, C.; Liu, X.; Dong, C.; Qin, Q. Bayesian Fusion of Multi-Scale Detectors for Road Extraction from SAR Images. *Int. J. Geo-Inform.* **2017**, *6*, 26. [CrossRef]
6. Tupin, F.; Maitre, H. Detection of linear features in SAR images: Application to road network extraction. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 434–453. [CrossRef]
7. Henry, C.; Azimi, S.M.; Merkle, N. Road Segmentation in SAR Satellite Images with Deep Fully Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
8. Wei, X.; Lv, X.; Zhang, K. Road Extraction in SAR Images Using Ordinal Regression and Road-Topology Loss. *Remote Sens.* **2021**, *13*, 2080. [CrossRef]
9. Jiang, M.; Miao, Z.; Gamba, P.; Yong, B. Application of Multitemporal InSAR Covariance and Information Fusion to Robust Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2017**, *99*, 3611–3622. [CrossRef]
10. Negri, M.; Gamba, P.; Lisini, G.; Tupin, F. Junction-aware extraction and regularization of urban road networks in high-resolution SAR images. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2962–2971. [CrossRef]
11. Tupin, F.; Houshmand, B.; Datcu, M. Road detection in dense urban areas using SAR imagery and the usefulness of multiple views. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2405–2414. [CrossRef]
12. He, C.; Bo, S.; Zhang, Y.; Xu, X.; Liao, M.S. Road extraction for SAR imagery based on the combination of beamlet and a selected kernel. In Proceedings of the Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014.
13. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]

14. Yang, X.; Li, X.; Ye, Y.; Lau, R.Y.K.; Zhang, X.; Huang, X. Road Detection and Centerline Extraction Via Deep Recurrent Convolutional Neural Network U-Net. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7209–7220. [CrossRef]
15. Wei, Y.; Zhang, K.; Ji, S. Simultaneous Road Surface and Centerline Extraction From Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8919–8931. [CrossRef]
16. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
17. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018.
18. Han, X.; Lu, J.; Zhao, C.; Li, H. Fully Convolutional Neural Networks for Road Detection with Multiple Cues Integration. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
20. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Trans. Geosci. Remote Sens.* **2018**, *15*, 749–753. [CrossRef]
21. Sun, T.; Chen, Z.; Yang, W.; Wang, Y. Stacked U-Nets with Multi-Output for Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
22. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogram. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
23. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
24. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–1924. [CrossRef]
25. Xu, Y.; Chen, H.; Du, C.; Li, J. MSACon: Mining Spatial Attention-Based Contextual Information for Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2021**, *99*, 1–17. [CrossRef]
26. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes From High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2043–2056. [CrossRef]
27. Chanussot, J.; Mauris, G.; Lambert, P. Fuzzy fusion techniques for linear features detection in multitemporal SAR images. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1292–1305. [CrossRef]
28. Głowacki, P.; Pinheiro, M.A.; Mosinska, A.; Türetken, E.; Lebrecht, D.; Sznitman, R.; Holtmaat, A.; Kybic, J.; Fua, P. Reconstructing Evolving Tree Structures in Time Lapse Sequences by Enforcing Time-Consistency. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 755–761. [CrossRef]
29. Ma, X.; Wang, C.; Yin, Z.; Wu, P. SAR Image Despeckling by Noisy Reference-Based Deep Learning Method. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8807–8818. [CrossRef]
30. Lu, X.; Zhong, Y.; Zhao, J. Multi-Scale Enhanced Deep Network for Road Detection. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2019), Yokohama, Japan, 28 July–2 August 2019; pp. 3947–3950. [CrossRef]
31. Dai, J.; Du, Y.; Zhu, T.; Wang, Y.; Gao, L. Multiscale Residual Convolution Neural Network and Sector Descriptor-Based Road Detection Method. *IEEE Access* **2019**, *7*, 173377–173392. [CrossRef]
32. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.; Paluri, M. Improved Road Connectivity by Joint Learning of Orientation and Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10377–10385. [CrossRef]
33. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef] [PubMed]
34. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T.S. CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 603–612. [CrossRef]
35. Chen, L.C.; Yi, Y.; Jiang, W.; Wei, X.; Yuille, A.L. Attention to Scale: Scale-Aware Semantic Image Segmentation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
36. Heipke, C.; Mayer, H.; Wiedemann, C.; Jamet, O. Evaluation of Automatic Road Extraction. *Int. Arch. Photogram. Remote Sens.* **1997**, *32*, 151–160.
37. Mnih, V.; Hinton, G. Learning to Label Aerial Images from Noisy Data. In Proceedings of the International Conference on Machine Learning, Edinburgh, UK, 26 June–1 July 2012.
38. Etten, A.V.; Lindenbaum, D.; Bacastow, T.M. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv* **2018**, arXiv:1807.01232v3.
39. Biagioni, J.; Eriksson, J. Inferring Road Maps from Global Positioning System Traces: Survey and Comparative Evaluation. *Transp. Res. Rec. J. Transp. Res. Board* **2014**, *2291*, 61–71. [CrossRef]