An Ensemble Wasserstein Generative Adversarial Network Method for Road Extraction From High Resolution Remote Sensing Images in Rural Areas

CHUAN YANG1,2 AND ZHENGHONG WANG3
1China Highway Engineering Consultants Corporation, Beijing 100089, China
2China Highway Engineering Consultants Corporation Data Company Ltd., Beijing 100089, China
3Department of Computer Science, University of Auckland, Auckland 1010, New Zealand
Corresponding author: Zhenghong Wang (zwan320@aucklanduni.ac.nz)

This work was supported by the Grand Special of High resolution On Earth Observation: Application demonstration system of high resolution remote sensing and transportation under Grant 07-Y30B03-9001-19/21.

ABSTRACT Road extraction from high resolution remote sensing (HR-RS) images is an important yet challenging computer vision task. In this study, we propose an ensemble Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) method called E-WGAN-GP for road extraction from HR-RS images in rural areas. The WGAN-GP model modifies the standard GANs with Wasserstein distance and gradient penalty. We add a spatial penalty term in the loss function of the WGAN-GP model to solve the class imbalance problem typically in road extraction. Parameter experiments are undertaken to determine the best spatial penalty and the weight term in the new loss function based on GaoFen-2 dataset. In addition, we execute an ensemble strategy in which we first train two WGAN-GP models using the U-Net and BiSeNet as generator respectively, and then intersect their inferred outputs to yield better road vectors. We train our new model with GaoFen-2 HR-RS images in rural areas from China and also the DeepGlobe Road Extraction dataset. Compared with the U-Net, BiSeNet, D-LinkNet and WGAN-GP methods without ensemble, our new method makes a good trade-off between precision and recall with $F_1$-score = 0.85 and IoU = 0.73.

INDEX TERMS Deep Learning; Road Extraction; High Resolution Remote Sensing; Wasserstein GAN; Ensemble Learning

I. INTRODUCTION

Up-to-date road networks are crucial for vehicle navigation, urban-rural planning, disaster relief and traffic management [1], [2]. However, manual road extraction is costly and slow, especially for rural areas which are weakly reachable and have limited access to GPS data unlike urban areas and highways [3], [4]. Instead, automatic road extraction from high resolution remote sensing (HR-RS) images have attracted much attention, since HR-RS images such as GanFen-2 (GF-2) have the advantages of wide imaging coverage, frequent revisit and sufficient spectral and spatial information of land covers [5], [6]. Most researchers regard road extraction from HR-RS images as a pixel-wise segmentation task which classifies each pixel into road or non-road according to geometric, photometric and texture features of roads (e.g. [2], [7], [8]). Thus, the performance of road extraction highly depends on the chosen semantic segmentation methods.

In the last decade, the deep convolutional neural network (CNN) has been introduced to the road extraction due to its impressive performances in computer vision [9]–[11]. CNN is a powerful feature extractor for road extraction which can automatically learn low to high level semantic features with consecutive convolutional operations. Based on CNN, full convolutional network (FCN) and its variants use encoder-decoder structures with skip connections to fuse multi-scale and multi-level features and have achieved state-of-the-art accuracy in road extraction tasks [8], [12]. Nevertheless, although numerous methods have been proposed over the decades, road extraction from HR-RS images is still far from large-scale practical application. The spectral heterogeneity in HR-RS images [6], [13] and the occlusions by trees, shadows and other non-road objects on roads are the
key factors that make the segmented road networks messy and unclean [14], [15]. A more general and robust road segmentation method is needed to improve the performance of road extraction from HR-RS images.

Recently, the generative adversarial networks (GAN) has taken over the deep learning community by storm due to their unparalleled flexibility and accuracy in image segmentation tasks [16]. It consists of a generator G and a discriminator D where G generates images that are indistinguishable from the ground truth and D distinguishes between the ground truth and the generated images. Few works have used standard GANs in road extraction from HR-RS images and achieved good performances [4], [7], [15], [17], [18]. However, these standard GANs may suffer from vanishing gradients, mode collapse, and unstable training [19], [20]. To solve the problems, [4] recently introduces the Wasserstein GAN with gradient penalty (WGAN-GP) model into road extraction. The WGAN-GP modifies the standard GANs using Wasserstein distance and gradient penalty which is more stable and easier to train [21]. However, his work neglects the class imbalance problem since the roads are spatially sparse in HR-RS images. In addition, his work only uses the U-Net [12] as G which can be improved by other state-of-the-art segmentation models such as BiSeNet [22]. And to further improve the overall performance in CNN methods, creating ensembles of multimodels may be an effective way [40]–[42].

In summary, the main contributions of this study are as follows:

1. A new WGAN-GP-based method named E-WGAN-GP is introduced into road extraction from HR-RS images in rural areas. The E-WGAN-GP method trains two WGAN-GP models with U-Net and BiSeNet as G, respectively. Then use an ensemble strategy of intersection calculation to improve the overall performance.

2. A spatial penalty term is added in the loss function of the WGAN-GP model to solve the class imbalance issue typically in road extraction.

3. Parameter experiments are undertaken to determine the best spatial penalty and the weight term of the E-WGAN-GP method in GF-2 dataset.

4. The E-WGAN-GP method outperforms the state-of-the-art method: U-Net, D-LinkNet, BiSeNet and WGAN-GP without ensemble in terms of $F_1$-score and IoU in both GF-2 and DeepGlobe datasets.

II. RELATED WORKS

There are intensive studies of road extraction from HR-RS images. Most of them use classification-based methods which can be further divided into supervised and unsupervised [23]. For supervised classification, [2] propose an artificial neural network (ANN) method with a Restricted Boltzmann Machines (RBM) initialization procedure to segment road. [24] use an object-oriented method based on support vector machine (SVM) to extract road centerline. [25] compare Markov random fields (MRF) with a hybrid model of the SVM and Fuzzy C-Mean (FCM) in road extraction. For unsupervised classification, [26] apply K-means clustering to extract road areas followed by mathematical morphology to remove non-road areas. [27] present a semi-automatic approach to delineate road networks incorporating the geodesic and mean shift methods. [28] introduce probabilistic and graph theory to infer road networks. Other classic methods represented by the snake and conditional random field (CRF) models are also adopted in road extraction. For example, [29] evolve a snake to capture road networks by minimizing an energy function which specifies geometric and appearance constraints of roads. [30] employ a high-order CRF to obtain the structures of road networks.

Inspired by the outstanding performances of deep learning in computer vision since 2012, CNN methods have been introduced into road extraction from HR-RS images. Compared with traditional road extraction methods which mostly utilize low level handcrafted features, CNN can automatically extract low to high level semantic features with consecutive convolutional operations [31], [32]. On the basis of CNN, FCN and its variants use encoder-decoder structures with skip connections to fuse multi-scale and multi-level features. In addition, they can output pixel-wise segmentation images with the same size as input images, which is suitable for end-to-end road extraction tasks [8], [12]. For example, [33] refine the U-Net with residual units to improve its performance. [34] propose the D-LinkNet which combines the LinkNet architecture with dilated convolutional layers. Other CNN methods are also adopted in road extraction. For example, [35] build the cascaded CNN (CasNet) to address road segmentation and centerline extraction tasks. [14] create the RoadTracer to get road network graphs following an iterative search procedure guided by a decision function based on CNN. [36] introduce DeepLab to detect roads with the MobileNet-v2 as the backbone. However, despite numerous studies, road extraction from HR-RS images is still far from large-scale practical application. The spectral heterogeneity in HR-RS images [6], [13] (e.g. large intra-class or small inter-class spectral variability) and the presence of occlusions (e.g. shadows and trees) often makes segmented road networks unclean and noisy [13]–[15]. A more general and robust road segmentation method is still needed to improve the performance of road extraction from HR-RS images.

The revolutionary technique in computer vision GAN is promising to fix these problems [16]. In a GAN we train two different neural networks called generator G and discriminator D, each one adversarial to the other. GAN-based methods regard road extraction as a task of image-to-image translation and work the same way as a conditional GAN (cGAN) [17], [37]. Few works have adopted the cGAN in road extraction from HR-RS images and made some progress. For example, [15] present a two-stage framework to extract roads, in which two GANs are used to detect roads and intersections followed by a smoothing-based optimization algorithm. [7] propose an improved GAN using the U-Net as G and suggest a simple loss function with an L2 loss and a cGAN loss. [18] create a multi-supervised GAN with two
D to infer road networks with improved topology. However, these methods are too complex either in architecture (e.g. two GANs in [15]) or loss function (e.g. four loss functions in [18]). Moreover, standard GANs struggle in vanishing gradients, mode collapse, and unstable training which are not easy to train [19], [20]. Recently, [4] introduces a WGAN-GP method to translate HR-RS images to a cleaner road network. The WGAN-GP model is proposed by [21] in 2017. It replaces the KL divergence in standard GANs with Wasserstein distance and employ a gradient penalty to enforce a Lipschitz constraint on the loss function. The WGAN-GP model has more stable training and outperforms other methods in road extraction [4]. Our study will further refine the WGAN-GP model to improve its performance in road extraction from HR-RS images in rural areas, which will be presented in the following sections.

III. METHODOLOGY

A. MODEL ARCHITECTURE

The architecture of WGAN-GP model follows the pix2pix [37], in which two adversarial models (a generator G and a discriminator D) are trained simultaneously [37]. G is given the image x in addition to a random noise z and aims to produce the realistic image G(x) similar to the label y, while D has to decide whether the image belong to y or obtained from G(x). The detailed architecture of WGAN-GP model is shown in Figure 1.

Since G is flexible with any segmentation models, we choose the widely used U-Net and the state-of-the-art BiSeNet as G respectively in this study. The U-Net is characterized by a symmetric encoder-decoder pathway with 4 down-sampling modules and 4 up-sampling modules (Figure 2a) [12]. The down-sampling module consists of repeated operations of two 3 × 3 convolutions, each followed by a batch normalization (BN), a rectified linear unit (ReLU) and a 2 × 2 max pooling operation. The up-sampling module contains a 2 × 2 up-convolution, a concatenation with the corresponding feature map from the decoder pathway at the same level, and two 3 × 3 convolutions, each followed by a BN and a ReLU. The BiSeNet is composed of a Spatial Path with a small stride to preserve the spatial information, and a Context Path with a fast down-sampling strategy (e.g. ResNet 101) to obtain sufficient receptive field (Figure 2b) [22]. To refine the output feature of each stage in the Context Path, the BiSeNet proposes an Attention Refinement Module (ARM) which employs global average pooling to capture global context and computes an attention vector to guide the feature learning. To fuse the output features of the two paths, the BiSeNet proposes a Feature Fusion Module (FFM) which employs concatenate operation and a weight vector to re-weight the features. In order to restrict our attention to the structure in local image patches, we use PatchGAN [37] discriminator following pix2pix with 5 down-sampling modules as D in the WGAN-GP model which classifies whether each N × N patch in an image is real or fake and average the classification outputs of each patch to obtain the final results (Figure 2c).

FIGURE 1. An overview of the architecture of WGAN-GP model. G produces the realistic image G(x) similar to the label y and is flexible with any segmentation models. D decides whether the image belongs to y or obtained from G(x). In order to restrict our attention to the structure in local image patches, we use PatchGAN as D.

FIGURE 2. Architectures of G and D in the WGAN-GP model: (a) Architecture of the U-Net in G (b) Architecture of the BiSeNet in G (c) Architecture of the PatchGAN in D. Each block in the architecture represents the feature map after down-sampling operations in which the length of the block represents the spatial resolution and the thickness represents the number of channels.

B. LOSS FUNCTION

The total loss function L_{total}(G,D) of our WGAN-GP model mixes the cGAN loss with the pixel-wise loss [37], which can be expressed as follows:

\[ \mathcal{L}_{cGAN}(G, D) + \beta \mathcal{L}_{L1}(G), \]

where \( \mathcal{L}_{cGAN}(G, D) \) is the standard cGAN loss, \( \mathcal{L}_{L1}(G) \) is the pixel-wise loss, \( \beta \) is the weight to balance the two losses.

The \( \mathcal{L}_{L1}(G) \) is used to evaluate the pixel-wise quality of the image generated by G, which encourages less blurring than the L2 loss [36]. The \( \mathcal{L}_{L1}(G) \) can be written as:

\[ E_{x,z,y} [\omega \| G(x, z) - y \|_1], \]

where \( x \) denotes the distribution of image x, \( z \) denotes the distribution of random noise, \( y \) denotes the distribution of label image y, and \( E_{x,y,z} \) is the expected value of the pixel-wise loss. Since roads are spatially sparse and only account for a small proportion of an image as shown in Figure 4, the class imbalance problem cannot be ignored [39], [40]. Here we add a spatial penalty term \( \omega \) to conserve the impact of the class imbalance. The \( \omega \) will be set to 1 if the pixel is the background and to \( \alpha \) (\( \alpha > 1 \)) if the pixel is the road. Accordingly, there will be a larger punishment if the road is
misclassified as the background:
\[
\omega = \begin{cases} 
\alpha & \text{if pixel } j \text{ is road} \\
1 & \text{if pixel } j \text{ is background}.
\end{cases}
\]

The \( L_{\text{GAN}}(G,D) \) is the standard cGAN loss and can be expressed as:
\[
E_{x,y} [\log D(x, y)] + E_{x,z} [\log (1 - (D(x, G(x, z)))]
\]
where \( G \) tries to minimize this loss while \( D \) tries to maximize it. \( E_{x,y} \) denotes the expected value over all real data instances, and \( E_{x,z} \) denotes the expected value over all generated fake instances \( G(x, z) \). However, [19] argued that the JS divergence used in standard GANs fails to provide a usable gradient when two distributions are disjoint, leading to the difficulty of training. They suggest replacing the JS divergence with the Wasserstein distance, which is continuous and differentiable almost everywhere.

Another [20]. In contrast to the JS divergence, the Wasserstein \( \lambda \)-distance is expressed as:
\[
\min_{G} \max_{D} \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z} [\log (1 - (D(x, G(x, z)))]
\]
\( \lambda \) is the penalty coefficient, \( \nabla G(\tilde{x}) \) is the gradient of the distribution \( \tilde{x} \), and \( E_{\tilde{x}} \) is the expected value of the gradient penalty term.

C. OPTIMIZATION PROCEDURE

The training dataset including image \( x \) and label \( y \) is fed to the model. The overall training objective aims at optimizing the total loss function \( L_{\text{total}}(G,D) \) to yield the final model \( G^* \), which can be expressed as follows:
\[
G^* = \arg \min_{G} \max_{D} L_{\text{GAN}}(G,D) + \beta L_{\text{L1}}(G).
\]

Specifically, an alternate optimization procedure is adopted. We first optimize \( D \) by fixing \( G \) (Equation 7) and then optimize \( G \) by fixing \( D \) (Equation 8), which is presented as follows:
\[
\max_{D} L_{\text{total}}(G,D) = E_{x,y} [\log D(x, y)] - E_{x,z} [\log (1 - D(x, G(x, z)))] + \lambda E_{\tilde{x}} [||\nabla G(\tilde{x})||_2 - 1].
\]
\[
\min_{G} L_{\text{total}}(G,D) = -E_{x,z} [D(x, G(x, z))] + E_{x,z,y} [\omega ||G(x, z) - y||_1].
\]

To accelerate the optimization speed, the mini-batch Adam is used. Repeat the training process till the training loss converges. A pseudocode for the optimization procedure is presented in Table 1.

D. ENSEMBLE STRATEGY

Due to the spectral heterogeneity in HR-RS images and the occlusions on roads, mistakes of road extraction are inevitable even for the state-of-the-art models. Creating ensembles of multi-models is proved to be an effective way of improving accuracies in CNN methods, because models trained with different architectures or hyper-parameters may have complementary information [40]–[42].

In this study, the ensemble strategy is carried out as follows: in the training stage, two WGAN-GP models in which \( G \) are tuned as U-Net and BiSeNet are trained respectively with the same training dataset and the same hyper-parameters. Here we do not train more models by tuning the hyper-parameters since too many models will affect the inference time. In the inference stage, we use the trained models to infer testing images and yield different sets of segmentation outputs. Since the intersection calculation of vector objects is more time-saving than fusing pixel values in segmentation images, we then apply a vectorization to the segmentation outputs and calculate the intersected vectors. Last we revert the intersected vectors to segmentation images and apply the skeleton thinning method followed by another vectorization to yield the final road vectors. Our ensemble strategy can take advantage of the best of different segmentation models and get road vectors with high precision as shown in Figure 5. The overall ensemble strategy is shown in Figure 3.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASETS

We use GF-2 images in this study. Launched in 2014, GF-2 satellite is the first civil optical remote sensing satellite in
China with a spatial resolution of 0.8 m. GF-2 is an ideal HR-RS data source for road extraction because of its high resolution, wide imaging coverage, frequent revisit and high image quality [6], [38].

We select over 30 scenes of GF-2 images in Fujian and Hainan Provinces from China covering various types of rural areas. The scenes are all in the visible spectrum with RGB bands. The pre-processing method contains radiometric calibration, atmospheric correction, and orthorectification. The original scenes are cropped to $512 \times 512$ pixels and ended up with 40000 images. Data augmentation methods including rotation and flipping are also applied to expand the dataset. We randomly split our dataset into a training dataset with 36000 images and a validation dataset with 4000 images. For testing, we choose extra three typical scenes of rural areas in Fujian and Hainan Provinces covering mountainous, plain and sub-urban areas, each with a size of $60000 \times 60000$ pixels.

Each image $x$ has an accompanying binary label $y$, indicating whether a pixel in the aerial image belongs to road (denoted as 1) or non-road (denoted as 0) as shown in Figure 4. To generate the label image, we follow the centerline-based approach to label road vectors manually in ArcGIS software and then rasterize them with a line width of 5 pixels [7], [39].

To verify the effectiveness of our method, we also conduct experiments on the famous DeepGlobe Road Extraction Challenge 2018 dataset [44]. We only use the training dataset of DeepGlobe dataset since the dataset has not published label images for the validation and testing dataset. The training dataset consists of a total of 6,226 HR-RS images with a size of $1024 \times 1024$ pixels. We randomly split our dataset into a training dataset with 5500 images, a validation dataset with 500 images and a testing dataset with 300 images. The spatial resolution of the dataset is 0.5 m which is higher than the GF-2 images.

**B. EXPERIMENTAL SETTINGS**

We implement our E-WGAN-GP method with the PyTorch framework and train them using NVIDIA Titan RTX GPUs with HR-RS images and paired label images from the training dataset. The spatial penalty term $\alpha$ in $L_{L1}(G)$ and the weight term $\beta$ in $L_{total}(G,D)$ are set to 100 and 10 according to our parameter-sensitivity experiment. The penalty coefficient $\lambda$ is set to 10 following [21]. For optimization, we use mini-batch Adam to train our model with $\beta_1 = 0.3$, $\beta_2 = 0.99$, weight decay $= 5e^{-5}$. The initial learning rate is set to 0.0005 and halves every 50 epochs. The max epoch is 300 and the batch size is 20.

**C. COMPARISON METHODS AND EVALUATION METRICS**

Our new method is compared with the other four methods: the U-Net [33], the BiSeNet, the D-LinkNet [34] and the WGAN-GP method using BiSeNet as $G$.

Traditional segmentation-based metrics such as precision, recall and IoU fail to be evaluation metrics in this study, since a slight error in the road width will heavily penalize the metrics while a small gap in an inferred road will not [43]. In this study, since our final outputs are road vectors, we suggest using vector-based precision, recall, $F_1$-score and IoU to evaluate the performance of our method. Precision and recall show the correctness and completeness of the inferred road vector, respectively. $F_1$-score considers the balance between the precision and recall and reaches its best value at 1. IoU represents the intersection of the prediction and ground truth over their region. Precision (P), recall (R), $F_1$-score ($F_1$) and IoU can be expressed as follows:

\[
P = \frac{TP}{TP + FP},
\]

\[
R = \frac{TP}{TP + FN},
\]

\[
F_1 = 2 \times \frac{P \times R}{P + R},
\]

\[
IoU = \frac{TP}{TP + FP + FN},
\]

where $TP$, $FN$ and $FP$ denote the true positive, false negative, and false positive respectively.

**D. EVALUATION ON GF-2**

We use hit/miss images to show our experimental results which superpose the extracted and labelled road vectors upon the original images to detect the hit and miss areas (Figure 5). We choose three typical scenes from rural areas to demonstrate the experimental results. Scene 1 represents a mountainous area where roads are distributed along valleys. Scene 2 displays a plain area with a river crossing by. Scene 3 is a sub-urban area with a huge body of water and many residences.

In the hit/miss images, yellow lines denote the roads which are extracted correctly ($TP$), red lines denote the roads which are missed ($FP$), and blue lines denote the roads which do not exist but are extracted incorrectly ($FN$). The E-WGAN-GP,
WGAN-GP (BiSeNet) and U-Net have much more yellow lines and fewer red lines than other methods suggesting they are superior in terms of R. The E-WGAN-GP and the D-LinkNet show much fewer blue lines than other methods indicating they outperforms others in terms of P. Accordingly, to consider both the R and P, the E-WGAN-GP method yields the best results in the hit/miss images.

**TABLE 2.** Quantitative comparison of different methods on the GF-2 testing dataset.

| Method         | P   | R   | F₁  | IoU  |
|----------------|-----|-----|-----|------|
| U-Net          | 0.41| 0.87| 0.55| 0.41 |
| BiSeNet        | 0.81| 0.76| 0.78| 0.66 |
| D-LinkNet      | 0.90| 0.59| 0.71| 0.60 |
| WGAN-GP (BiSeNet)| 0.83| 0.82| 0.82| 0.71 |
| E-WGAN-GP      | 0.86| 0.85| 0.85| 0.73 |

Table 2 shows a quantitative comparison of five methods, including P, R, F₁ and IoU. The E-WGAN-GP and U-Net give the highest R (0.85-0.87), and the E-WGAN-GP and D-LinkNet yield the highest P (0.86-0.90). As a result, our E-WGAN-GP makes the best trade-off between P and R with the highest F₁ (0.85) and IoU (0.73) and is more reliable in large-scale practical applications.

To get a deep understanding of the errors our E-WGAN-GP makes in GF-2 dataset, Figure 6 shows one result from Scene 2 with paired original image and hit/miss image. Typical FNs are in areas where roads are thin and narrow or hard to be distinguished from the background due to shadows and trees (purple box). Possible ways to tackle these problems include using different road widths in dataset generation [7] and improve the post-processing method to the segmentation images. Typical FPs are caused by either narrow roads (blue box) or interior roads in residences (orange box). One possible way to tackle this problem is adding more corresponding negative samples to the training dataset.

**TABLE 3.** Quantitative comparison with different parameters for the E-WGAN-GP in GF-2 DATASET.

| Parameter | P   | R   | F₁  | IoU  |
|-----------|-----|-----|-----|------|
| No α, β=10| 0.81| 0.41| 0.55| 0.43 |
| α=5, β=10 | 0.86| 0.55| 0.67| 0.57 |
| α=100, β=1| 0.81| 0.82| 0.81| 0.68 |
| α=100, β=10| 0.86| 0.85| 0.85| 0.73 |
| α=200, β=10| 0.72| 0.87| 0.79| 0.69 |
| α=100, β=100| 0.75| 0.86| 0.80| 0.71 |

**E. PARAMETER ANALYSIS**

Since the performance of the E-WGAN-GP method is highly parameter-sensitive in its loss function, we undertake experiments to determine the best parameters in GF-2 dataset. The parameters involved in our experiments are the spatial penalty term α in \( L_{\text{L1}}(G) \) and the weight term β in \( L_{\text{total}}(G,D) \) which are the most critical parameters during training. From Figure 7 and Table 3, we can observe that adding the spatial penalty term α in \( L_{\text{L1}}(G) \) will significantly improve the performance. In addition, when we increase α and β (α from 5 to 100 and β from 1 to 10), IoU increases from 0.57 to 0.74 as well. These results suggest that the spatial penalty term α may help us eliminate the class imbalance issue typically in road extraction task [39], [40].

However, when α and β become too large (α from 100 to 200 and β from 10 to 100), IoU decreases from 0.74 to 0.69 instead. This may be caused by that many non-road objects are “punished” to be roads by the large α and β. We suggest the proper range of the α and β to be 50-150 and 5-50, respectively. In this study, we set the α and β as 100 and 10 in our study.

**F. EVALUATION ON DEEPGLOBE**

To verify the effectiveness of our method, we also compare the performance of our method on the public DeepGlobe Road Extraction Challenge 2018 dataset. The comparison
Table 4. Quantitative comparison of different methods on the DeepGlobe testing dataset.

| Method      | P    | R    | F1   | IoU  |
|-------------|------|------|------|------|
| D-LinkNet   | 0.89 | 0.70 | 0.77 | 0.67 |
| E-WGAN-GP   | 0.79 | 0.85 | 0.82 | 0.73 |

The quantitative comparison including P, R, F1 and IoU is displayed in Table 4. Our method has higher averaged R, F1 and IoU compared with the D-LinkNet, suggesting the effectiveness of our method in road extraction from HR-RS data.

V. CONCLUSION

In this study, we present an E-WGAN-GP method for road extraction from HR-RS images in rural areas. The WGAN-GP model used in our method is a modification of standard GANs with Wasserstein distance and gradient penalty. To solve the class imbalance problem specially in road extraction, we add a spatial penalty term in the loss function. Parameter experiments are undertaken to determine the best parameters of the new loss function based on GaoFen-2 dataset. We also follow the ensemble strategy first to train different WGAN-GP models with different G (U-Net and BiSeNet) and then to ensemble their results to yield better road vectors. The E-WGAN-GP method outperforms the state-of-the-art method: U-Net, D-LinkNet, BiSeNet and WGAN-GP without ensemble in terms of F1-score and IoU in both GF-2 and DeepGlobe datasets.

REFERENCES

[1] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks,” ISPRS J. Photogramm. Remote Sens., vol. 130, pp. 139–149, Aug. 2017.

[2] V. Mnih and G. E. Hinton, “Learning to detect roads in high-resolution aerial images,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2010, pp. 210–223.

[3] S. He, F. Bastani, S. Abbar, M. Aliizadeh, H. Balakrishnan, S. Chawla, and S. Madden, “RoadRunner: Improving the precision of road network inference from GPS trajectories,” in Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., Nov. 2018, pp. 3–12.

[4] E. Park, “Refining inferred road maps using GANs,” Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2019.

[5] A. Buslaev, S. Seferbekov, V. Iglovikov, and A. Shvets, “Fully convolutional network for automatic road extraction from satellite imagery,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 207–210.

[6] X.-Y. Tong, Q. Lu, G.-S. Xia, and L. Zhang, “Large-scale land cover classification in Gaofen-2 satellite imagery,” in Proc. IGARSS-IEEE Int. Geosci. Remote Sens. Symp., Jul. 2018, pp. 3599–3602.

[7] X. Zhang, X. Han, C. Li, X. Tang, H. Zhou, and L. Xiao, “Aerial image road extraction based on an improved generative adversarial network,” Remote Sens., vol. 11, no. 8, p. 930, Apr. 2019.

[8] 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.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.

[11] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” 2019, arXiv:1905.11946. [Online]. Available: http://arxiv.org/abs/1905.11946.

[12] 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.

[13] W. Xia, Y. Z. Zhang, J. Liu, L. Luo, and K. Yang, “Road extraction from high resolution image with deep convolution network—A case study of GF-2 image.” Multidisciplinary Digita1. Publishing Inst. Proc., vol. 2, no. 7, p. 325, 2018.

[14] F. Bastani, S. He, S. Abbar, M. Aliizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, “RoadTracer: Automatic extraction of road vectors from aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4720–4728.

[15] D. Costea, A. Marcu, M. Leordeanu, and E. Slusanschi, “Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 2100–2109.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[17] Q. Shi, X. Liu, and X. Li, “Road detection from remote sensing images by generative adversarial networks,” IEEE Access, vol. 6, pp. 25486–25494, 2018.

[18] Y. Zhang, Z. Xiong, Y. Yang, C. Wang, J. Li, and X. Li, “Toplogy-aware road network extraction via multi-supervised generative adversarial networks,” Remote Sens., vol. 11, no. 9, p. 1017, Apr. 2019.

[19] M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” 2017, arXiv:1701.04862. [Online]. Available: http://arxiv.org/abs/1701.04862.

[20] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” 2017, arXiv:1701.07875. [Online]. Available: http://arxiv.org/abs/1701.07875
[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein GANs,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5767–5777.

[22] C. Yu, J. Wang, C. Peng, C. Gao, G. Yuv, and N. Sang, “Bisenet: Bilateral segmentation network for real-time semantic segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 325–341.

[23] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, and P. Eklund, “A review of road extraction from remote sensing images,” J. Traffic Transp. Eng. (English Ed.), vol. 3, no. 3, pp. 271–282, 2016.

[24] X. Huang and L. Zhang, “Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines,” Int. J. Remote Sens., vol. 30, no. 8, pp. 1977–1987, Apr. 2009.

[25] D. M. Zhu, X. Wen, and C. L. Ling, “Road extraction based on the algorithms of MRF and hybrid model of SVM and FCM,” in Proc. Int. Symp. Image Data Fusion, Aug. 2011, pp. 1–4.

[26] R. Maurya, P. R. Gupta, and A. S. Shukla, “Road extraction using K-means clustering and morphological operations,” in Proc. Int. Conf. Image Inf. Process., Nov. 2011, pp. 1–6.

[27] Z. Miao, B. Wang, W. Shi, and H. Zhang, “A semi-automatic method for road centerline extraction from VHR images,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 11, pp. 1856–1860, Nov. 2014.

[28] C. Unsalan and B. Sirmacek, “Road network detection using probabilistic and graph theoretical methods,” IEEE Trans. Geosci. Remote Sens., vol. 50, no. 11, pp. 4441–4453, Nov. 2012.

[29] I. Laptev, H. Mayer, T. Lindeberg, W. Eckstein, C. Steger, and A. Baumgartner, “Automatic extraction of roads from aerial images based on scale space and snakes,” Mach. Vis. Appl., vol. 12, no. 1, pp. 23–31, Jul. 2000.

[30] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, “Road networks as collections of minimum cost paths,” ISPRS J. Photogramm. Remote Sens., vol. 108, pp. 128–137, Oct. 2015.

[31] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[32] W. Zhao, S. Du, and W. J. Emery, “Object-based convolutional neural network for high-resolution imagery classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

[33] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual U-Net,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 749–753, May 2018.

[34] L. Zhou, C. Zhang, and M. Wu, “D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 182–186.

[35] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, “Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[36] Y. Nachmany and H. Alemohammad, “Detecting roads from satellite imagery in the developing world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2019, pp. 83–89.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[38] Q. Wu, R. Zhong, W. Zhao, K. Song, and L. Du, “Land-cover classification using GF-2 images and airborne lidar data based on random forest,” Int. J. Remote Sens., vol. 40, nos. 5–6, pp. 2410–2426, May 2019.

[39] M. Yuan, Z. Liu, F. Wang, and F. Jin, “Rethinking labelling in road segmentation,” Int. J. Remote Sens., vol. 40, no. 22, pp. 8359–8378, Nov. 2019.

[40] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, “Fully convolutional network-based ensemble method for road extraction from aerial images,” IEEE Geosci. Remote Sens. Lett., early access, Dec. 4, 2019, doi: 10.1109/LGRS.2019.2953523.

[41] C. Ju, A. Bibaut, and M. van der Laan, “The relative performance of ensemble methods with deep convolutional neural networks for image classification,” J. Appl. Statist., vol. 45, no. 15, pp. 2800–2818, Nov. 2018.

[42] F. Perez, S. Avila, and E. Valle, “Solo or ensemble? Choosing a CNN architecture for melanoma classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 2775–2783.

[43] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, “SpaceNet: A remote sensing dataset and challenge series,” 2018, arXiv:1807.01232. [Online]. Available: http://arxiv.org/abs/1807.01232

[44] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “DeepGlobe 2018: A challenge to parse the Earth through satellite images,” 2018, arXiv:1805.06561. [Online]. Available: http://arxiv.org/abs/1805.06561

CHUAN YANG received the B.S. degree in geology from the China University of Geosciences, Wuhan, China, in 2013, and the Ph.D. degree in mineralogy, petrology, mineral geology from Peking University, Peking, China, in 2018. His research interests are in deep learning, remote sensing, and geology.

ZHENGHONG WANG received the B.S. degree in computer science from the University of Auckland, New Zealand, in 2018, where he is currently pursuing the M.S. degree in information technology. His research interests are in data mining and machine learning.