IMPROVING SEMANTIC SEGMENTATION OF HIGH-RESOLUTION REMOTE SENSING IMAGES USING WASSERSTEIN GENERATIVE ADVERSARIAL NETWORK

H.R. Hosseinpour [1*], F. Samadzadegan [1], F. Dadrass Javan [1,2], S. Motayyeb [1]

[1] School of Surveying and Geospatial Information Engineering, College of Engineering, University of Tehran, Tehran, Iran
[2] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 NB Enschede, the Netherlands

Commission IV, WG IV/3

ABSTRACT:

Semantic segmentation of remote sensing images with high spatial resolution has many applications in a wide range of problems in this field. In recent years, the use of advanced techniques based on fully convolutional neural networks have achieved high and impressive accuracies. However, the labels of different classes are estimated independently in this method. In general, the segmentation effect is too coarse to take the relationship between pixels into account. On the other hand, due to the use of convolution filters and limitations of calculations, the field of view information of these filters will be limited in deep layers. In this study, a method based on generative adversarial network (GAN) is proposed to strengthen spatial vicinity in the output segmentation map. The segmentation model receive assistance from the GAN model in the form of a higher order potential loss. Furthermore, for better stability and performance in model training the Wasserstein GAN is used for optimization of the model. We successfully show an increase in semantic segmentation accuracy using the challenging ISPRS Vaihingen benchmark dataset.

KEY WORDS: Semantic Segmentation, Deep Learning, Wasserstein GAN, Generative Adversarial Network

1. INTRODUCTION

One of the fundamental and difficult issues in remote sensing and photogrammetry is the semantic segmentation of high-resolution areal and satellite imagery, which tries to give labels to each pixel in an image (Hua et al., 2021; Zhang et al., 2016). Semantic segmentation has gained increasing attention from researchers in recent years due to its numerous applications in a variety of higher-level remote sensing tasks (Yuan et al., 2021), including urban feature extraction such as building area (Hosseinpour et al., 2022a), change detection (Zheng et al., 2021), etc. Accurate segmentation results, both for localization and classification, are still exceedingly difficult to obtain (Asgari Taghanaki et al., 2021). The difficulties in this assignment are the complicated background, the significant diversity in appearance, the numerous perspectives and poses of various objects, etc.

A new method for semantic segmentation of remote sensing images has recently been proposed by Deep Convolutional Neural Network (DCNN) (Yuan et al., 2021). The effective processing of remote sensing images can be done using these DCNN-based segmentation approaches. Since the input image pixels in this method are predicted independently, this causes a lot of inconsistency in the semantic segmentation of images. Post-processing methods have been proposed to overcome this problem. However, many of these post-processing methods have limitations such as computation (Papadomanolaki et al., 2018; Vakalopoulou et al., 2015). Another problem of DCNN is that in these networks, due to the limitations of the calculations, the dimensions of the calculated features for the input image in the deeper layers have a lower resolution. This is based on down-sampling operations. Fully connected networks (FCNs) (Long et al., 2014) were made available to address this issue. In FCN, encoder-decoder networks were employed. In this network, the global information is discarded while completely connected layers are removed to obtain correct spatial information. After presenting the FCN method, various deep learning models consisting of encoder and decoder parts have been presented to perform the image segmentation process. One of these successful architecture models is SegNet (Badrinarayanan et al., 2017). In the encoder part, the VGG network is used, and in the last layers of this network, due to the down-sampling operation, the spatial resolution of features is reduced, and instead, the semantic concepts of features are increased. The most important advantage of the SegNet method is the storage of the index address in the polling operation in the encoder part, which is finally used to recover the features in the decoder part. In research (Ronneberger et al., 2015) the famous UNet model is presented. In this research, similar to the SegNet method, two stages of encoder and decoder are used. However, the features generated in the encoder part are concatenated with the decoder cross section. In the research (Chen et al., 2018) the Deeplab model is presented. Atrous convolution method is used in this model. By using atrous convolution method, in which the dimensions of the convolution filter are changed, different resolution of feature can be achieved. In recent years, semi-supervised methods based on adversarial models have been used for segmentation. For example, in research (Xu and Wang, 2021) a semi-supervised method was presented in which a generator was used to generate training samples and a discriminator was used to label complications. In the research (Li et al., 2022), a method similar to the previous method was used in the segmentation of remote sensing images.

In this paper in contrast to the semi-supervised method, the higher order discrepancies between the ground truth maps and the segmentation prediction maps are then attempted to be corrected by jointly utilizing the GAN's generator and discriminator (Tolstikhin et al., 2014). We use the architecture of pix2pix (Isola et al., 2017) as the GAN model

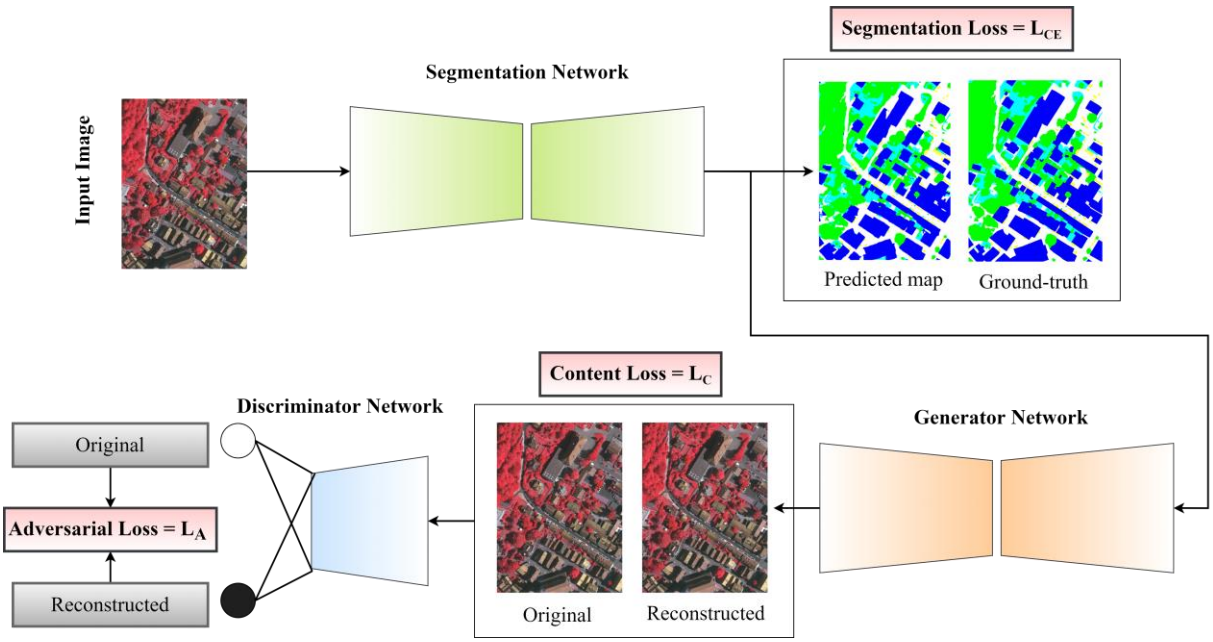


Figure 1. Illustration of the proposed semantic segmentation network.

and the architecture of our previous segmentation network (Hosseinpour et al., 2022b) is employed as the basic segmentation model. To optimize the proposed segmentation network, a composite objective function is suggested. The content loss provided by the generator, the adversarial loss taken from the discriminator, and the traditional multi-class cross-entropy loss are all incorporated in the proposed objective function. The segmentation model is then optimized using the suggested scheme. Additionally, the training of our model incorporates the Wasserstein GAN (WGAN) optimization approach that was suggested by (Arjovsky et al., 2017). The outcomes demonstrate that we achieve greater performance and stability.

2. METHODOLOGY

As shown in Figure 1, three fundamental components make up the architecture described in this study. The segmentation model, which function is to create the prediction maps, is included in the first part. The second part includes generator model, is responsible for reconstructing the original image from the outputs of the segmentation model's last layer. The third part is called discriminator and its task is to compare the reconstructed image with the real image. At first, an initial training is done on GAN using real data. Finally, by using the pre-trained model as an additional loss function, it is used to optimize the classification model. In this research, WGAN is specifically used to train the GAN network.

2.1 Segmentation model

In this research, the segmentation model presented in the (Hosseinpour et al., 2022b) has been used. This model, which is designed based on the famous U-Net architecture, includes two encoder and decoder parts. In the encoder part, resnet50 architecture has been used, in which all the final fully connected layers have been removed. The function of the encoder part is to extract the feature space. based on this part during the training of the network, the spatial resolution of the feature decreases, but the number of channels increases. In the decoder part, an attempt is made to increase the spatial resolution of the features through up-sampling operations. In this part, the RUM+ architecture is used after concatenation the features of the decoder and encoder layers.

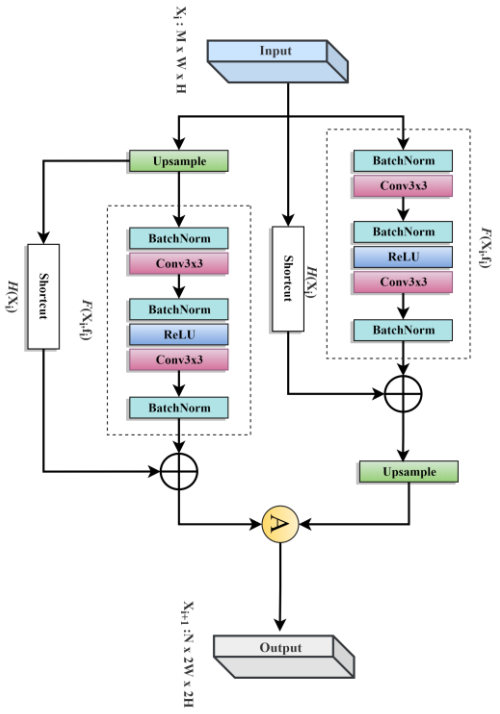


Figure 2. The proposed RUM+ structure used in the decoder part of the segmentation network.

Figure 2 shows the proposed RUM+ structure in the decoder section. In this structure, at first, the incoming features is sent to two separate streams. In each of these two streams, the structure of the residual networks is used. However, in one flow, first the up-sampling operation is done and it is sent to the residual network, and in the other flow, after the residual network of the features space, it is transferred to the up-sampling part. Finally, the output of two streams is merged

with each other through averaging. The structure of the residual part in both streams is based on the proposed structure presented in.

2.2 GAN model

The GAN network used in this research consists of two parts: generator and discriminator, which compete in a game. The loss function of the current GAN determines how each network plays the game. The structure of generator and discriminator in this research is in accordance with the research pix2pix (Isola et al., 2017; R et al., 2019) and is indicated by G and D respectively.

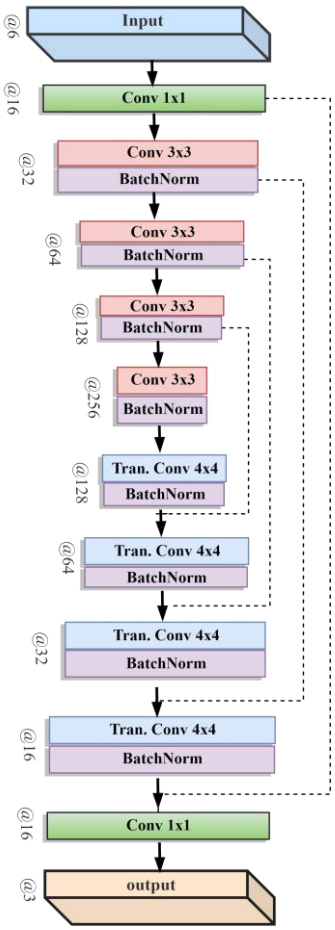


Figure 3. The structure of the generator.

Figure 3 shows the structure of generator network architecture. In this structure, which is similar to the structure of the UNet architecture (Ronneberger et al., 2015), short connections between convolution blocks are used in encoder and encoder networks. The input feature space contains a 6-channel tensor, which is basically the output of the segmentation network (in this case, 5 channels are related to the classes of objects in the image and one channel is related to the background of the image.). The output from the generator network is a reconstructed image with 3 channels. The architectural structure of the discriminator is shown in Figure 4. This architecture consists of 4 convolution layers along with Batch-normalization layer.

The value of the loss function is the goal of both the generator and the discriminator. The G wants to minimize the loss function and the D wants to maximize it. In other words, in the min-max optimization process, D can supervise G. Network G learns how to create real images like the original image, meanwhile, network D tries to discriminate between the image obtained from network G and the original image. The following relationship can be used to define how to train a GAN architecture:

$$\min_G \max_D L(D,G) = \mathbb{E}_{I\sim P_r(I)}\left[\log D(I)\right] + \mathbb{E}_{I\sim P_g(I^p)}\left[\log(1-D(I^p))\right] \quad (1)$$

Where I represent image from the data distribution over original images P_r, I^p denotes the reconstructed image from G, and Pg represent the generator's distribution over original images.

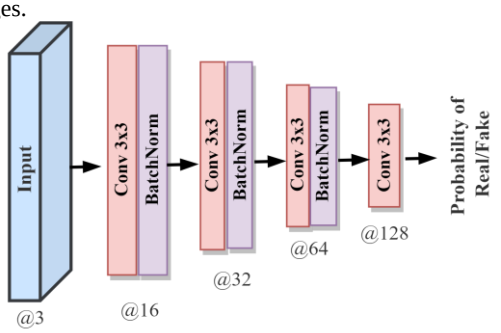


Figure 4. The structure of the discriminator

2.3 Loss function

In this study, the final hybrid loss function L consist of three main loss functions, the segmentation loss function L_{CE}, the content loss function L_C and the adversarial loss L_A:

$$L = L_{CE} + \alpha L_C + \beta L_A \quad (2)$$

Where α and β represent the weight parameters. For L_{CE}, cross entropy loss for multi-class segmentation is used.

$$L_{CE} = -\sum_{i}^{C} g_i \log(f(p_i)) \quad (3)$$

Where p_i and g_i are the score from the activation function (f(.)) (i.e., softmax or sigmoid) and ground-truth for each class i in C. L_c determine the quality of the reconstructed image I^p generated by generator network and according to the Wasserstein GAN (Arjovsky et al., 2017), L_A can reflect the quality of the reconstructed images. These two-loss function are formulated as:

$$L_C = \mathbb{E}_{I\sim P_r(I)}\left[\|I - G(I)\|_1\right] \quad (4)$$

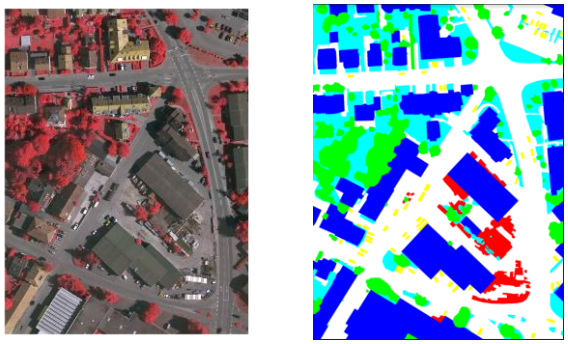
$$L_A = -\mathbb{E}_{I\sim P_g(I^p)}\left[D(G(I^p))\right] \quad (5)$$

Before using equation (2) to train the segmentation network, the proposed GAN model should be trained first. The following loss function (equation (6)) is used to train this network. In the next step, the segmentation network is trained and the trained coefficients are considered fixed in the GAN network. Basically, the GAN network is used as a support network to increase the accuracy of the segmentation network.

$$L_A = -\mathbb{E}_{I\sim P_g(I^p)}\left[D(G(I^p))\right] \quad (6)$$

3. EXPERIMENTAL RESULTS

In this research, to evaluate the effectiveness and capabilities of the proposed method, an experiment was conducted on the ISPRS Vaihingen dataset[1]. The images of Vaihingen dataset are cropped from a large orthophoto image associated with a part of the city of Vaihingen in Germany. The average size of the images in this collection is 2000 × 2500 pixels. The spatial resolution of these images is 0.09 meters. Each image is presented in three near-infrared, red, and green bands. Ground-truth images include 5 classes of impervious surface, buildings, low vegetation, trees, cars, and one class of other background features. In this dataset, based on the information provided by the data provider, 16 images were proposed as training data to train and evaluate of the model, and the rest of the images were used as test data. Figure 5 shows an example of the image used in the training phase of the proposed network along with the ground-truth data.



(a) (b)

Figure 5. Example of the ISPRS Vaihingen image used for training with corresponding ground-truth.

In this research, in order to evaluate the proposed method, two well-known criteria of overall accuracy (OA) and the average of intersection over union (IoU) have been used. In order to determine these two criteria, the confusion matrix has been used. For this purpose, in order to determine the value of OA, which is also expressed as pixel accuracy, the ratio of the total number of pixels that are correctly classified divided to the total number of pixels in the evaluation dataset. Equation (7) can be used to calculate this criterion. In this regard, True-positive (TP) is equivalent to the total number of pixels that are correctly classified in the desired class. True-negative (TN) represents the total number of background pixels that are correctly classified based on the ground-truth data. False-negative (FN) expresses the number of pixels from the background that are wrongly classified in the desired class, and false-positive (FP) expresses the number of pixels from the desired class that are wrongly placed in the background class.

$$OA_c = \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \quad (7)$$

The IoU metric is calculated to express the percentage of overlap between the image pixels and the ground-truth pixels, and it can be calculated by measuring the number of common pixels between the target and prediction masks divided by the total number of pixels in both masks. The IoU values for each class are calculated separately and finally the average value for all classes is presented as the result of the segmentation method. The equation (8) is used to calculate the value of IOU in each class. The value of index c in equations (7) and (8) expresses the class of complications for calculating metric values.

$$IOU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (8)$$

The proposed method is implemented in PyTorch framework (Paszke et al., 2019). As mentioned in the previous section, the training of the proposed method is done in two parts. First, the generative adversarial network was trained to learn the distribution of ground-truth data. For this purpose, taking into account the limitations of the GPU hardware, the images with dimensions of 512 × 512 and batch size equal to 4 was set. Adam method (Kingma and Ba, 2014) is used for optimization. In addition, before using the images in the training process, in order to argument data and prevent the problem of overfitting, pre-processing operations were performed on the images, including horizontal and vertical flipping, rotating and resizing the images. The number of epochs for training the GAN network was considered equal to 60.

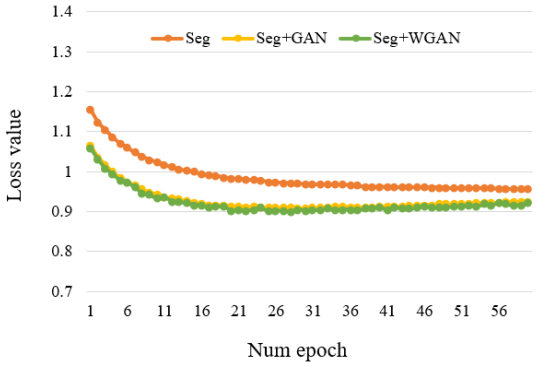


Figure 6. Training loss values of proposed methods on the Vaihingen dataset.

In the next step, the training of the segmentation network was integrated by considering the trained GAN model. Therefore, three methods can be considered. The method in which the training of the network is done without considering GAN ($Model_{Seg}$) and the method in which the segmentation network is trained with the original GAN method ($Model_{Seg+GAN}$). The third mode in model training can be considered the simultaneous training of the segmentation network along with the Wasserstein GAN model ($Model_{Seg+WGAN}$), which in this case uses the proposed loss function provided in equation (2). In this case, the weight parameters in equation (2) are considered equal to 0.1. In Figure 6, the convergence process of the loss function for training the model is mentioned in three methods and shown for 60 iteration loops.

After training the segmentation network using the 3 methods mentioned above, to test the results, the model is first called and then its parameters are initialized using the weight parameters of the trained model. In the next step, the test images similar to the training step are cut to the size of 512 × 512, so that the images overlap each other by 50%. The final segmentation results are thresholded after averaging between the results. The visual results of this implementation are shown in Figure 7. In this figure, two test images are used. The Vaihingen dataset is considered a challenging dataset where there are many differences between the classes in scale, dimensions and texture. In this dataset, the impervious

[1] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

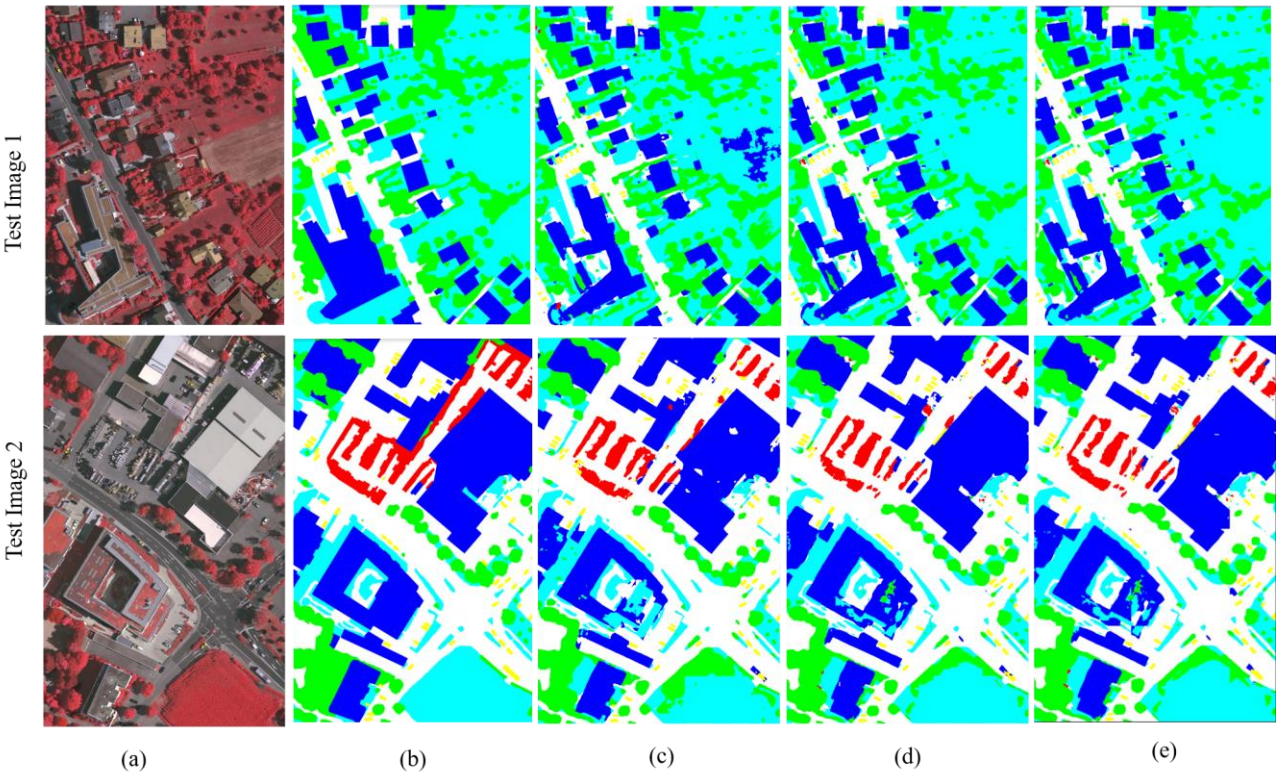


Figure 7. Semantic segmentation result for two sample images from the ISPRS Vaihingen dataset. (a) Input image, (b) Ground-truth, (c) $Model_{Seg}$, (d) $Model_{Seg+GAN}$, (e) $Model_{Seg+WGAN}$

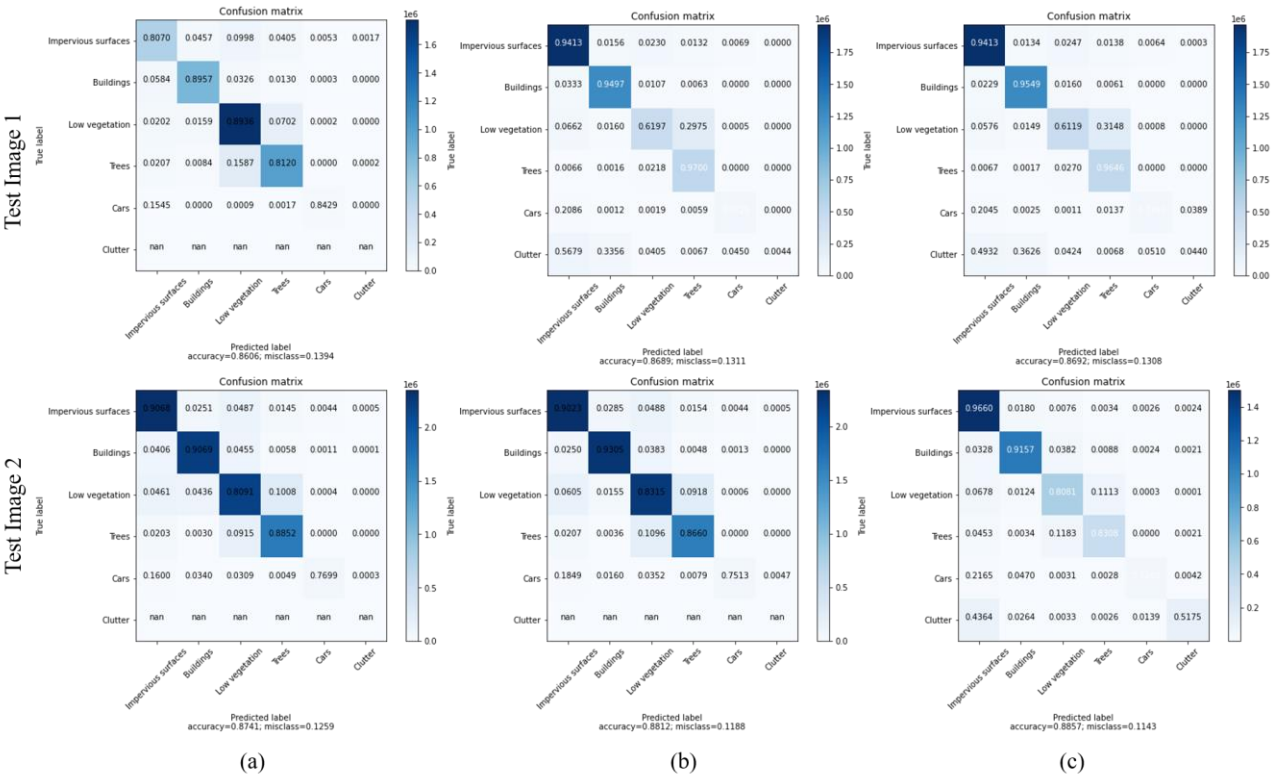


Figure 8. Confusion matrix of the proposed semantic segmentation network. (a) $Model_{Seg}$, (b) $Model_{Seg+GAN}$, (c) $Model_{Seg+GAN}$

surface class is shown in white, building features in blue, trees in green, vegetation features in cyan, cars in yellow, and background objects in red. The results of the implementation of methods $Model_{Seg+GAN}$ and $Model_{Seg+WGAN}$ have shown the superiority of the proposed model in semantic segmentation. As shown in test image 1, the building class is detected with less noise than in methods $Model_{Seg}$. In other image classes, we see the improvement of the visual results, especially in the border areas of complications. This shows that the GAN model has the ability to effectively detect higher order

potentials and the classes in the images are correctly identified. In Figure 8, the confusion matrix associated with two test images is shown, where we see the increase of the OA criterion in the semantic segmentation of complications. In addition, the use of method WGAN has improved the results compared to method GAN. which shows the high ability of the model in solving the problem of semantic classification of remote sensing images.

Methods	Overall Accuracy (%)	mIoU (%)
$Model_{Seg}$	87.20	64.28
$Model_{Seg+GAN}$	87.94	65.32
$Model_{Seg+WGAN}$	88.06	65.44

Table 1. Result on ISPRS Vaihingen test images.

Table 1 shows the quantitative results of three methods for all the test images in the Vaihingen dataset. According to the third column of this table, the average value of the IOU of methods A and B has a significant improvement compared to the classification model. Therefore, it can be concluded that the GAN method as an additional loss function has a high ability to improve semantic segmentation results.

4. CONCLUSION

Semantic segmentation of remote sensing images with high spatial resolution is considered as an important problem with many applications in the field of remote sensing. In this research, a new framework has been proposed in which the integration of a segmentation model and a GAN model is used for semantic segmentation in order to improve the results. For this purpose, the GAN model is first trained, in which the high-order inconsistencies between the image and the ground-truth data are measured through the loss function. In the next step, the learned GAN network is considered as a helper loss function to train and adjust the parameters of the semantic segmentation model. In this research, experiments were conducted to check the performance of the proposed model on the ISPRS Vaihingen dataset. The results show the improvement of the performance of the proposed model.

REFERENCES

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. https://doi.org/10.48550/arxiv.1701.07875

Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G., 2021. Deep semantic segmentation of natural and medical images: a review. Artif. Intell. Rev. 54, 137–178. https://doi.org/10.1007/s10462-020-09854-1

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022a. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 184, 96–115. https://doi.org/10.1016/J.ISPRSJPRS.2021.12.007

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022b. A Novel Boundary Loss Function in Deep Convolutional Networks to Improve the Buildings Extraction From High-Resolution Remote Sensing Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 4437–4454. https://doi.org/10.1109/JSTARS.2022.3178470

Hua, Y., Marcos, Di., Mou, L., Zhu, X.X., Tuia, D., 2021. Semantic Segmentation of Remote Sensing Images with Sparse Annotations. IEEE Geosci. Remote Sens. Lett. 19. https://doi.org/10.1109/lgrs.2021.3051053

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-January, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization 1–15.

Li, J., Liao, Y., Zhang, J., Zeng, D., Qian, X., 2022. Semi-Supervised DEGAN for Optical High-Resolution Remote Sensing Image Scene Classification. Remote Sens. 14, 4418. https://doi.org/10.3390/rs14174418

Long, J., Shelhamer, E., Darrell, T., 2014. Fully Convolutional Networks for Semantic Segmentation. 2019 IEEE/CVF Int. Conf. Comput. Vis. Work. 847–856. https://doi.org/10.1109/ICCVW.2019.00113

Papadomanolaki, M., Vakalopoulou, M., Paragios, N., Karantzalos, K., 2018. Stacked Encoder-Decoders for Accurate Semantic Segmentation of Very High Resolution Satellite Datasets. IGARSS 2018 - 2018 IEEE Int. Geosci. Remote Sens. Symp. 6927–6930. https://doi.org/10.1109/igarss.2018.8519113

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv.

R, J.V.C.I., Zhu, X., Zhang, Xinming, Zhang, Xiao-yu, Xue, Z., Wang, L., 2019. A novel framework for semantic segmentation with generative adversarial network q. J. Vis. Commun. Image Represent. 58, 532–543. https://doi.org/10.1016/j.jvcir.2018.11.020

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Comput. Vis. Pattern Recognit. 1–8.

Tolstikhin, I., Bousquet, O., Schölkopf, B., Thierbach, K., Bazin, P.L., de Back, W., Gavriilidis, F., Kirilina, E., Jäger, C., Morawski, M., Geyer, S., Weiskopf, N., Scherf, N., Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B.,

Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., Musk, E., Neuralink, Hjortsø, M.A., Wolenski, P., Ruder, S., Grathwohl, W., Chen, R.T.Q., Bettencourt, J., Sutskever, I., Duvenaud, D., Doersch, C., 2014. Generative Adversarial Networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11046 LNCS, 1–9. https://doi.org/10.48550/arxiv.1406.2661

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. Int. Geosci. Remote Sens. Symp. 2015-Novem, 1873–1876. https://doi.org/10.1109/IGARSS.2015.7326158

Xu, D., Wang, Z., 2021. Semi-supervised semantic segmentation using an improved generative adversarial network. J. Intell. Fuzzy Syst. 40, 9709–9719. https://doi.org/10.3233/JIFS-202220

Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. 169, 114417. https://doi.org/10.1016/J.ESWA.2020.114417

Zhang, Liangpei, Zhang, Lefei, Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4, 22–40. https://doi.org/10.1109/MGRS.2016.2540798

Zheng, Z., Wan, Y., Zhang, Y., Xiang, S., Peng, D., Zhang, B., 2021. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 175, 247–267. https://doi.org/10.1016/j.isprsjprs.2021.03.005