Remote Sensing Image Segmentation based on Implicit 3D Scene Representation

Zipeng Qi, Zhengxia Zou, Hao Chen, and Zhenwei Shi⋆, *Member, IEEE*

***Abstract*—Remote sensing image segmentation, as a challenging but fundamental task, has drawn increasing attention in the remote sensing field. Recent advances in deep learning have greatly boosted research on this task. However, the existing deep learning-based segmentation methods heavily rely on a large amount of pixel-wise labeled training data, and the labeling process is time-consuming and labor-intensive. In this paper, we focus on the scenario that leverages the 3D structure of multi-view images and a limited number of annotations to generate accurate novel view segmentation. Under this scenario, we propose a novel method for remote sensing image segmentation based on implicit 3D scene representation, which generates arbitrary-view segmentation output from limited segmentation annotations. The proposed method employs a two-stage training strategy. In the first stage, we optimize the implicit neural representations of a 3D scene and encode their multi-view images into a neural radiance field. In the second stage, we transform the scene color attribute into semantic labels and propose a ray-convolution network to aggregate local 3D consistency cues across different locations. We also design a color-radiance network to help our method generalize to unseen views. Experiments on both synthetic and real-world data suggest that our method significantly outperforms deep convolutional networks (CNN)-based methods and other view synthesis-based methods. We also show that the proposed method can be applied as a novel data augmentation approach that benefits CNN-based segmentation methods.**

***Index Terms*—Remote sensing, Image segmentation, Implicit neural representations, Neural radiance field.**

I. Introduction

Remote sensing image segmentation, as a fundamental but very challenging task, aims at classifying the input remote sensing image pixel by pixel into different categories. Remote sensing image segmentation is widely used in various applications, including building detection [1], road detection [2], hyperspectral image classification [3] etc. With the rapid development of deep learning technology, the segmentation method based on convolution neural networks (CNN) has achieved excellent performance and has become a research hot spot. Despite the recent progress in this field, the training of CNN-based methods may heavily rely on a large amount of pixel-wise labeled data, while the labeling process is time-consuming and labor-intensive. Recent remote sensing image segmentation methods [4]–[8] place a greater emphasis on 2D texture features while disregarding 3D spatial information, which results in a scene segmentation with poor view consistency. To address the above problems, this paper studies an interesting question: is it possible to improve remote sensing image segmentation by leveraging 3D view consistency between different views, particularly, under the condition of a limited number of labels (e.g. about 5% of total annotations)? To answer this question, we propose a novel method for remote sensing image segmentation based on implicit 3D scene representation, which generates arbitrary-view segmentation output from limited segmentation annotations. Fig. 1 shows an overall idea of the proposed method.

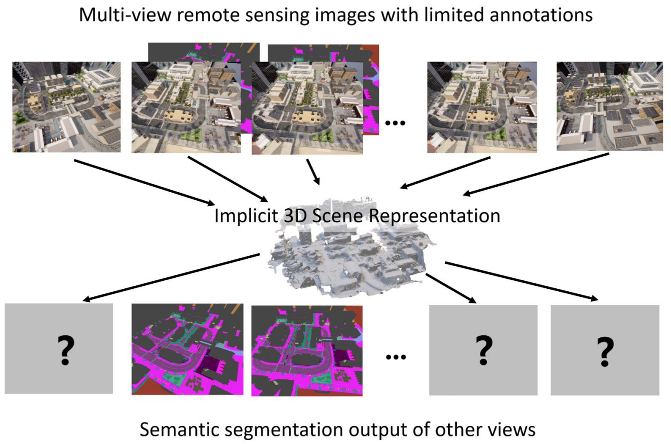

Multi-view remote sensing images with limited annotations
...
Implicit 3D Scene Representation
?
...
?
?
Semantic segmentation output of other views

Fig. 1. In this paper, we study an interesting question - given a set of multi-view remote sensing images, how do generate semantic segmentation outputs from sparse annotations by leveraging 3D structural consistency between different views? To answer this question, we propose a novel remote sensing segmentation method based on implicit 3D scene representation and shows superiority over state-of-the-art CNN-based methods.

Our approach can be viewed as an extension of 3D vision in the direction of 2D visual understanding. Recently, the novel view synthesis [9] and 3D scene representation [10], [11] have aroused great research interest in the field of vision and graphics. The representation of 3D scenes can be divided into two groups, explicit representations, and implicit representations. In explicit representations, such as mesh [12], voxel [13], and TSDF [14], the 3D spatial information can be conveniently accessed and edited. However, it also requires a lot of storage space, and it is difficult to give an accurate

The work was supported by the National Key Research and Development Program of China (Titled "Brain-inspired General Vision Models and Applications"), the National Natural Science Foundation of China under Grant 62125102, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Zipeng Qi, Hao Chen and Zhenwei Shi are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and with State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory.

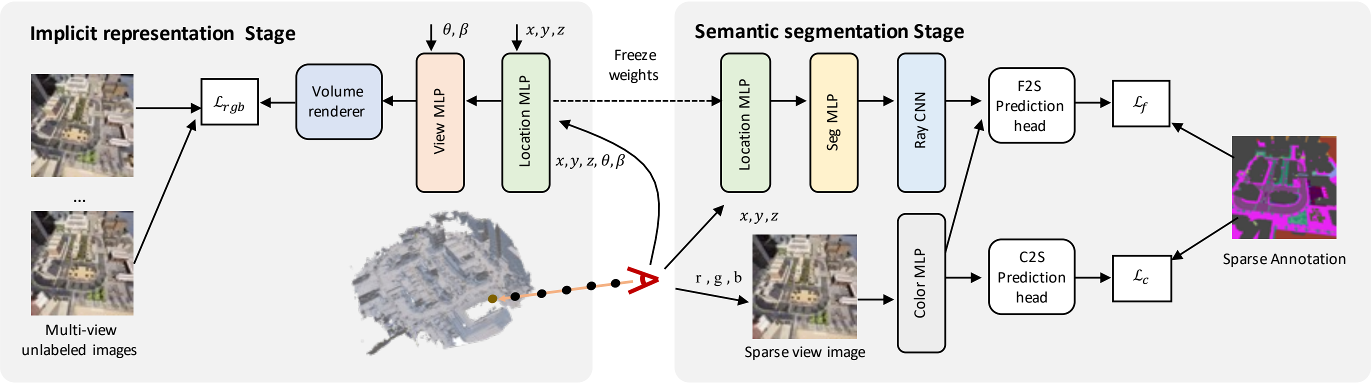


Fig. 2. An overview of our method. Our method consists of two processing stages: an implicit representation stage and a semantic segmentation stage. In the first stage, we utilize multi-view RGB images to extract implicit spatial information; in the second stage, we partially fix the parameters of the implicit neural network and propose a ray convolutional network and a color radiation network to better produce the scene's semantic attributes under sparse annotation.

representation of the details.

More recently, the implicit neural representation has shown powerful capabilities in implicitly encoding 3D scenes [10], [15], [16]. NeRF [15] is a high-profile work in this field that renders photo-realistic novel-view images using optimized density and color attributes. Specifically, it optimizes the weights of a neural network to represent the mapping between the spatial coordinates and corresponding 3D attributes. Compared with explicit representation, implicit representation is more flexible and shows greater potential. In this paper, we propose an implicit representation-based method for the scenario we focus on. The proposed method employs a two-stage training strategy that differs from the conventional CNN-based image segmentation paradigm to solve the scenario. In the first stage, the implicit 3D representation of a scene is encoded in the weights of Multilayer Perceptron (MLP) by optimizing the density and color attributes of the spatial points in a continuous volumetric representation. In the second stage, with optimized density attributes and a few labels, color attributes are transformed into semantic attributes. Finally, we render the accurate segmentation results for each test view. Due to insufficient annotation, the features of some spatial locations cannot be fully optimized in the sparse view. To address this issue, the color-radiance network and ray-convolution network are proposed. The color-radiance network extracts 2D color features and fuses them with 3D spatial features to improve the accuracy of prediction results. The Ray-convolution integrates the 3D features of local neighborhood sampled points within a viewing ray to improve the network's representation of the semantic attributes.

To quantitatively evaluate our method, we build a multi-view aerial remote sensing dataset named Carla-MVS based on the well-known Carla simulation platform [17]. The proposed method outperforms other CNN-based state-of-the-art image segmentation methods under the condition of limited labeled data. We also conduct qualitative experiments on real-world remote sensing images from Google Earth, our results are more accurate and view-consistent. Furthermore, our method can also benefit downstream applications - we show that using our generated segmentation results as a new data augmentation strategy, the performance of CNN-based methods can be significantly improved.

II. Methodology

A. Overview

In this paper, a two-stage implicit neural field optimization method for remote sensing scene segmentation is proposed. An overall of our method is shown in Fig. 2. 1) In the implicit representation stage, we follow the NeRF [15] pipeline and feed the 3D coordinates and the view angle of the 3D locations to two multilayer perceptrons, i.e., a location-MLP and a view-MLP. In this way, the coordinate-dependent density attribute and view-dependent color attribute can be generated for each 3D location. 2) In the semantic segmentation stage, we effectively used the spatial information extracted from the above stage to generate segmentation results. Specifically, we freeze the weights of the Location-MLP so that the spatial information implied by the density attribute can be reused. A prediction head named Seg-MLP is then introduced to generate semantic features based on density features. Considering the insufficiency of the annotation of limited views, we also propose a ray-convolution network, where pixel features are extracted and fused with semantic features. This enables the model to properly employ spatial information and pixel information to produce a more accurate segmentation output.

B. Implicit representation stage

In the stage of implicit representation, we aim at constructing an implicit 3D representation of the target remote sensing scene. The stage-1 process is shown in the left part of Fig 2. Similar to NeRF [15], we sample spatial points along the ray formed by each pixel in the multi-view image and then estimate the attributes for each spatial point. We use spatial point coordinates (x, y, z) and ray angle (θ,β) as inputs for the NeRF MLPs Φ_{nerf}:

$$\sigma_i, c_i = \Phi_{nerf}(x, y, z, \theta, \beta), \tag{1}$$

where $\sigma_i \in R^1$ and $c_i \in R^3$ are the density and color attributes of the i_{th} 3D location, respectively. After that, the pixel color corresponding to the ray is rendered through discrete integration [15]:

$$\hat{C}(r) = \sum_{i=1}^{N} \exp\Big(- \sum_{j=1}^{i-1} \alpha_j \sigma_j \Big) \Big(1 - \exp\left(-\alpha_i \sigma_i\right) \Big) c_i, \quad (2)$$

where α_i is the distance between two adjacent sample points and N is the number of sample points along a ray. Finally, we refer to the corresponding pixel in the RGB image as the ground truth $C(r)$ to optimize the implicit neural field.

$$\mathcal{L}_{rgb} = \sum_{r \in R} \Big[\| \hat{C}(r) - C(r) \|_2^2 \Big]. \quad (3)$$

where R are the sampled rays within a training batch.

C. Semantic segmentation stage

After obtaining the implicit 3D representation of the target scene, in the semantic segmentation stage, we aim at generating the semantic segmentation output given any view image. Different from the first stage, we only feed the coordinates of the spatial points into MLPs in the above branch. Owning to the continuity of implicit neural representation and the accuracy of the density attribute, the semantic features of points within the public view region can be well optimized.

Considering the limited annotations, we design the color-radiance network and ray-convolution network as follows:

1) Ray-convolution network: In remote sensing scenes, we assume the semantic attribute of spatial points along a ray is ideally consistent. The ray-convolution network enhances the association between semantic attributes of neighborhood points along a ray. We choose the 1×3 convolution kernel to process the semantic attributes of adjacent sampling points in a ray rather than the typical $n \times n$ convolution kernel:

$$f_i = \Phi_{cnn}(f_{i-1}, f_i, f_{i+1}), \quad (4)$$

where the $f_i \in R^{1 \times 1 \times 128}$ is the feature of each spatial point after the ray-convolution network. We then use the following formulation to generate semantic features along the ray:

$$S_s(r) = \sum_{i=1}^{N} \exp\Big(- \sum_{j=1}^{i-1} \alpha_j \sigma_j \Big) \Big(1 - \exp\left(-\alpha_i \sigma_i\right) \Big) s_i, \quad (5)$$

where $s_i \in R^{1 \times 1 \times 128}$ is the semantic features of the i_{th} sampling point.

2) Color-radiance network: As an additional branch, the color-radiance network assists the model in predicting the semantic attributes of points. Like other MLP-based modules, the color-radiance network consists of a set of fully-connected layers and Relu layers. Differently, the input is the color vector (R, G, B) of the pixel corresponding to the ray, and the output $N(r) \in R^{1 \times 1 \times 128}$ as the overall features of the ray are to be concatenated with $S_s(r)$. The fused features are then fed into the F2S prediction head (F2S: Fused features to Segmentation) to get the final semantic results $\hat{S}_f(r) \in R^{1 \times L}$. The network at this stage are optimized by referring to the segmentation ground-truth $S(r)$.

$$\mathcal{L}_f = - \sum_{r \in R} \Big[\sum_{l=1}^{L} S^l(r) \log \hat{S}_f^l(r) \Big]. \quad (6)$$

where L is the number of classes and R are the sampled rays within a training batch. In addition, we feed $N(r)$ into another classification layer, C2S-MLP (C2S: Color to Segmentation) to obtain a semantic result $\hat{S}_c(r) \in R^{1 \times L}$ and apply the following loss to optimize the color-radiance network to ensure the accuracy of $N(r)$.

$$\mathcal{L}_c = - \sum_{r \in R} \Big[\sum_{l=1}^{L} S^l(r) \log \hat{S_c}^l(r) \Big] \quad (7)$$

D. Training details

Our proposed method is implemented with Pytorch and trained on an RTX3090 GPU. We use Adam optimization to optimize the networks, and all compared methods are retrained on our dataset. We set the learning rate to $5e^{-4}$ in the first stage and $1e^{-3}$ in the second stage. To make a fair comparison, we set the number of optimizations in the first stage is 200000, the same as seg-NeRF [18]. Similar to NeRF, we adopt two-stage sampling to discretize the scene. In the first stage, we evenly sample 64 points along a ray that passes through the center of the camera and each pixel in the input image. In the second stage, we further sample 192 points by importance sampling based on the first stage results to make the points more concentrated near the object.

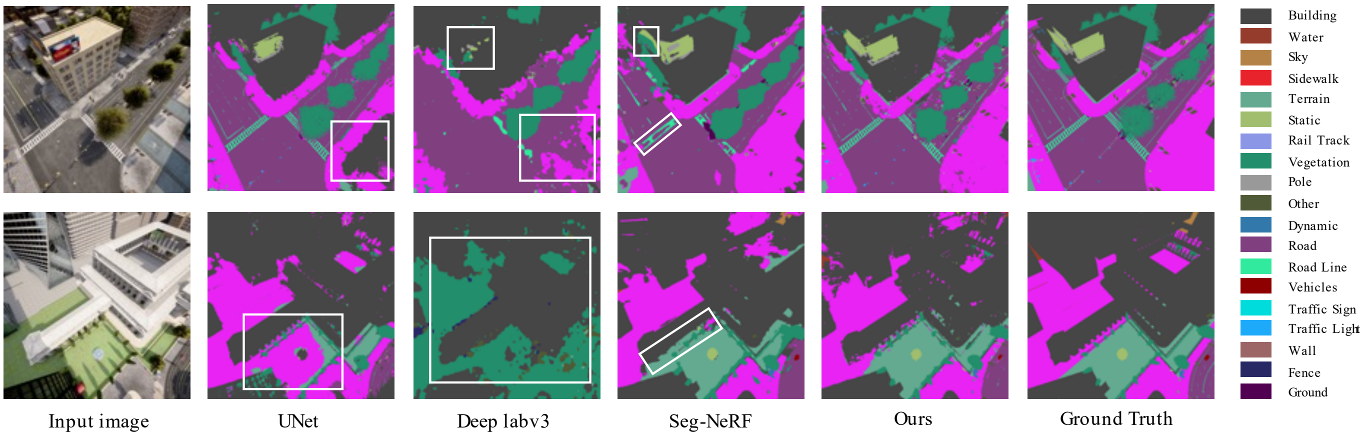
III. Experiment and Analysis

We compare three well-known CNN-based segmentation methods and a state-of-the-art NeRF-based segmentation method: Unet [19] is designed with a skip connection to effectively fuse the feature information between encoder and decoder to preserve more details. Chen designed the atrous convolution in Deeplab [7] that enlarges the receptive field of the convolution kernel without additional computation. Fu proposed Dual Attention Network (DANet) [8] to adaptively integrate local features and global dependencies. Seg-NeRF [18] is a recently proposed NeRF-based framework for semantic segmentation. We retrained all the compared methods using our dataset, all using single-view images as input to the network.

A. Dataset and Metrics

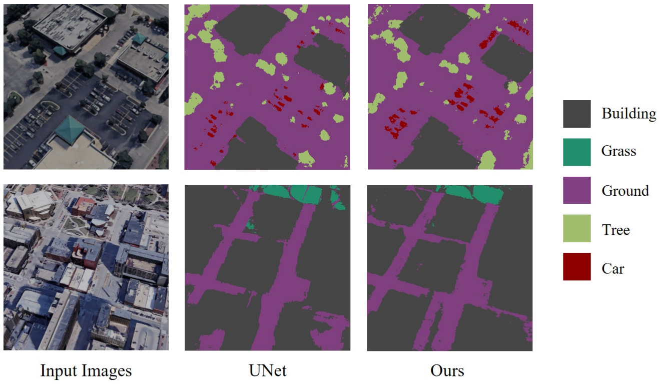
1) Dataset: We test on both synthetic and real-world data. The synthetic dataset Carla-MVS (Carla multi-view segmentation) consists of five subsets with images captured above the urban environment from the open-source CARLA program [17]. The statistics of the five subsets are shown in table I. For each scene, only 3% - 6% images are annotated. The real-world images are from Google Earth satellite images. We estimate camera parameters for each image in all datasets using COLMAP.

2) Metrics: Model performance evaluation is based on the similarity of the segmentation results to the Ground Truths. The current automatic evaluation metrics mainly include mIoU, Avg Acc (Class Average Accuracy), and Total Acc (Pixel Average Accuracy), which are used in [18]. For all the metrics, a higher score means a better result.



Input image UNet Deep labv3 Seg-NeRF Ours Ground Truth

Fig. 3. Image segmentation results on the Carla-MVS dataset. The white boxes mark the wrong-segmentation regions. Compared with CNN-based and NeRF-based methods, our method produces more complete and detailed results.



Input Images UNet Ours

Fig. 4. Image segmentation result on real-world data from Google Maps.

TABLE I
DATASET STATISTICS

	#views	#annoations	#classes	size (pixels)
subset #1	100	3	20	512 × 512
subset #2	100	4	18	512 × 512
subset #3	100	5	18	512 × 512
subset #4	80	5	19	512 × 512
subset #5	85	5	19	512 × 512

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS.

	Methods	mIOU	Total Acc	Avg Acc
subset #1	Unet [19]	23.92	74.51	33.28
	DANet [8]	9.73	65.41	12.00
	Deeplabv3 [7]	18.89	74.04	22.90
	Seg-NeRF [18]	55.73	93.87	62.09
	Ours	**58.03**	**94.33**	**64.94**
subset #2	Unet [19]	31.73	85.31	36.88
	DANet [8]	13.91	75.16	17.83
	Deeplabv3 [7]	16.64	75.26	21.68
	Seg-NeRF [18]	34.81	85.89	43.07
	Ours	**38.46**	**86.50**	**47.29**
subset #3	UNet [19]	38.72	90.07	43.07
	DANet [8]	15.79	75.15	19.76
	Deeplabv3 [7]	20.52	81.19	24.05
	Seg-NeRF [18]	41.85	**91.20**	48.78
	Ours	**43.53**	90.47	**49.41**
subset #4	UNet t [19]	41.94	91.85	44.79
	DANet [8]	26.78	80.00	31.07
	Deeplabv3 [7]	30.42	82.34	34.83
	Seg-NeRF [18]	57.24	95.29	61.33
	Ours	**59.14**	**95.72**	**66.02**
subset #5	UNet [19]	26.63	88.19	29.82
	DANet [8]	8.53	74.12	10.72
	Deeplabv3 [7]	11.90	77.64	14.61
	Seg-NeRF [18]	41.44	88.80	51.33
	Ours	**43.77**	**91.52**	**51.56**

B. Experiments and Results

1) Segmentation accuracy: Since CNN is sensitive to view changes and changes in the number of sample annotations, it makes it difficult for CNN-based methods to obtain accurate prediction results consistent with views. From Table II, we can see our method increases the accuracy of the results significantly compared to the CNN-based method, indicating that using implicit neural representation and 3D consistency helps overcome the aforementioned issues. By further incorporating the semantic features and color information, our model outperforms the seg-NeRF [18].

2) Visual comparison: From Fig. 3, we can see our method produces more complete and detailed results compared with CNN-based methods and seg-NeRF. We also validated our method on real-world data, as shown in Fig 4. Compared to UNet, our results are more consistent from various perspectives and have better visual effects.

3) Ablation study: We evaluate the effectiveness of different components of our method, including the color-radiance network and the ray-convolution network. The results are shown in Table III.

- Effectiveness of the ray-convolution network: From Table III, we see the ray-convolution network improves the average mIoU about by 3.1% compared to its baseline. This indicates integrating semantic features along a ray makes the rendering output more consistent with the ground truth.

TABLE III
RESULTS OF OUR ABLATION ON RAY-CONVOLUTION NETWORK AND COLOR-RADIANCE NETWORK. FOR BREVITY, WE USE R FOR THE RAY-CONVOLUTION AND C FOR THE COLOR-RADIANCE NETWORK.

Ablations			mIoU on Carla-MVS Dataset					
Our baseline	R	C	subset #1	subset #2	subset #3	subset #4	subset #5	Avg accuracy
✓			53.19	24.70	41.22	55.22	39.38	42.74 ± 10.99
✓	✓		55.70	35.53	41.03	55.84	41.08	45.84 ± 8.36
✓	✓	✓	58.03	38.46	43.53	59.14	43.77	**48.59** ± 8.39

- Effectiveness of the color-radiance network: The color-radiance network fuses color features and semantic features, which further improves the accuracy of the model by 2.7%, as shown in Table III. Fig. 3 also suggests our results are more detailed and complete.

4) Downstream validation: Our method can be also used as a novel approach for data augmentation. We take UNet [19] as a baseline and test whether our method can benefit segmentation. From Table IV, we see that UNet's performance on the Carla-MVS has been significantly improved with our method.

TABLE IV
SEGMENTATION ACCURACY IMPROVEMENT OF UNET [19] BY USING OUR METHOD AS A DATA AUGMENTATION STRATEGY.

	mIOU	Total Acc	Avg Acc
subset #1	57.85 (↑ 33.93)	94.30 (↑ 19.79)	64.64 (↑ 31.36)
subset #2	38.65 (↑ 6.92)	89.86 (↑ 4.55)	43.24 (↑ 6.36)
subset #3	43.48 (↑ 4.76)	90.93 (↑ 0.86)	48.57 (↑ 5.50)
subset #4	50.68 (↑ 8.78)	94.69 (↑ 2.85)	53.40 (↑ 8.61)
subset #5	42.94 (↑ 16.31)	91.41 (↑ 3.22)	48.69 (↑ 18.87)

IV. CONCLUSION

We propose a novel method for remote sensing image segmentation based on implicit neural representation. Under the spare annotations, we achieve more accurate and detailed results compared to CNN-based methods. We propose a ray-convolution network to integrate semantic features in a ray space and a color-radiance network to fuse the pixel color features. Experimental results show that our strategy effectively eliminates the model's sensitivity to view change under a limited number of labels, resulting in more accurate and view-consistent segmentation results. Moreover, our results can be viewed as a new data augmentation strategy that helps improve the performance of CNN-based remote sensing image segmentation methods.

REFERENCES

[1] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention unet for building segmentation in remote sensing images," *Science China Information Sciences*, vol. 63, no. 4, pp. 1–12, 2020.

[2] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, "Road segmentation for remote sensing images using adversarial spatial pyramid networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 4673–4688, 2020.

[3] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3232–3245, 2019.

[4] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[5] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network," *Remote sensing*, vol. 11, no. 15, p. 1774, 2019.

[6] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[9] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv preprint arXiv:2010.07492*, 2020.

[10] L. Shen, J. Pauly, and L. Xing, "Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[11] E. Dupont, H. Loya, M. Alizadeh, A. Goliński, Y. W. Teh, and A. Doucet, "Coin++: Data agnostic neural compression," *arXiv preprint arXiv:2201.12904*, 2022.

[12] Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao, "Meshnet: Mesh neural network for 3d shape representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8279–8286.

[13] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[14] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "Neuralrecon: Real-time coherent 3d reconstruction from monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 598–15 607.

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[16] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4799–4809, 2019.

[17] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[18] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 838–15 847.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.