Implicit Ray-Transformers for Multi-view Remote Sensing Image Segmentation

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

***Abstract*—The mainstream CNN-based remote sensing (RS) image semantic segmentation approaches typically rely on massively labeled training data. Such a paradigm struggles with the problem of RS multi-view scene segmentation with limited labeled views due to the lack of consideration of 3D information within the scene. In this paper, we propose "Implicit Ray-Transformer (IRT)" based on Implicit Neural Representation (INR) for RS scene semantic segmentation with sparse labels (5% of the images being labeled). We explore a new way of introducing the multi-view 3D structure priors to the task for accurate and view-consistent semantic segmentation. The proposed method includes a two-stage learning process. In the first stage, we optimize a neural field to encode the color and 3D structure of the remote sensing scene based on multi-view images. In the second stage, we design a Ray Transformer to leverage the relations between the neural field 3D features and 2D texture features for learning better semantic representations. Different from previous methods that only consider 3D priors or 2D features, we incorporate additional 2D texture information and 3D priors by broadcasting CNN features to different point features along the sampled ray. To verify the effectiveness of the proposed method, we construct a challenging dataset containing six synthetic sub-datasets collected from the Carla platform and three real sub-datasets from Google Maps. Experiments show that the proposed method outperforms the CNN-based methods and the state-of-the-art INR-based segmentation methods in quantitative and qualitative metrics. The ablation study shows that under a limited number of fully annotated images, the combination of the 3D structure priors and 2D texture can significantly improve the performance and effectively complete missing semantic information in novel views. Experiments also demonstrate the proposed method could yield geometry-consistent segmentation results against illumination changes and viewpoint changes. Our data and code will be public.**

***Index Terms*—Remote sensing, implicit neural representation, semantic segmentation, Transformer**

I. Introduction

Remote sensing image segmentation is a fundamental yet challenging task that has been widely applied in various fields, such as cloud detection [1], change detection [2], and land analysis [3]. The objective of remote sensing image segmentation is to produce pixel-wise labels for an image. Thanks to the advancement of imaging technology and satellite technology, it is now easy to acquire high-resolution multi-view RGB images.

The mainstream segmentation methods [4–6] benefit from deep convolutional neural networks (CNN), which can effectively learn and extract robust and discriminative features from the input images. However, deep CNN-based segmentation methods rely heavily on massive training data. As shown in Fig. 1(a), the performance of traditional CNN-based methods is sensitive to the number of annotations. Generating a large number of high-quality pixel-wise annotations consumes a great deal of time and effort. As for the task of semantic segmentation for a 3D scene given only limited annotated views, the CNN-based methods may overfit the views in the training data and generate poor results for the rest of the views. The key reason is that the 2D texture information or 2D context relationship is insufficient to identify similarly textured objects (Fig. 1(b)) in a 3D scene. Finally, the 3D context relationship of a scene is also crucial for semantic attribute prediction (Fig.1(c)). For example, the bridge is typically higher than the road, and the same object across different views usually has a similar texture. However, these properties have been rarely investigated in previous papers.

Considering the above challenges, this paper studies the task of multi-view remote sensing scene semantic segmentation under limited annotations, as shown in Fig. 2. We show that the 3D structure priors are crucial for this task. The representation of the 3D structure is a fundamental and long-studied problem in computer vision and graphics. There are many explicit representation-based 3D reconstruction methods [7–9] that extract 3D context information in different forms, including depth maps [10], meshes [11], and point clouds [12]. However, these methods typically require explicit supervision data which are hard to obtain and computationally intensive. Recently, an emerging research topic named "Implicit Neural Representation (INR)" has made rapid advances, particularly in the realm of novel view synthetic [13–16]. The INR provides a novel way to parameterize continuous signals driven by coordinates with neural networks. The INR-based novel view synthetic fits images from known viewpoints and utilizes the continuity of the spatially-varying scene properties (such as color and geometry) in high-dimensional space to render compelling photo-realistic novel view images. During this process, the geometry and color attributes are encoded into the weights of a neural network. Compared to explicit reconstruction, the INR is more flexible to optimize without

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under the Grants 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhengxia Zou (e-mail: zhengxiazou@buaa.edu.cn))*

Zipeng Qi, Hao Chen, Chenyang Liu, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, with the Beijing Key Laboratory of Digital Media, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

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, Shanghai 200232, China.

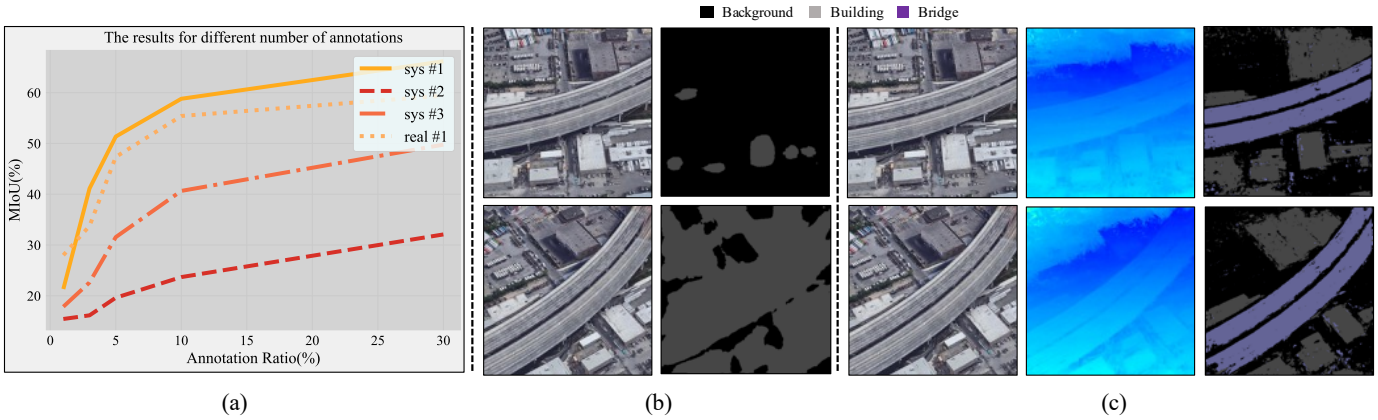

Fig. 1. (a): The image shows the performance of Unet with different annotation ratios; (b): The images show that it is difficult for the CNN-based methods to distinguish objects with similar textures under the limited number of annotations; (c): Our proposed method can use the 3D information of the scene to assist in distinguishing objects with similar textures, such as backgrounds and bridges. The middle images in (c) show the 3D information extracted by our method in the form of depth maps.

expensive supervision.

In this paper, inspired by the INR in novel view synthesis, we propose a new framework for multi-view remote sensing image segmentation under limited view annotations which utilizes an INR to exploit 3D structure priors. We also propose a new network architecture called "Ray-Transformer" that combines the 3D structure and 2D texture information from a set of 3D location points along the rays. We refer to our method as "Ray-Transformer" and refer to the task we study as Remote Sensing Scene Semantic Segmentation (R4S). Given a set of multi-view RGB images, the proposed method can generate accurate and semantically consistent novel view segmentation output even only trained with a limited number of labels (e.g., 4-6 labels per 100 images).

The proposed method has a two-stage learning process. We first optimize a color-INR of the target scene using multi-view RGB images, where the 3D context information is encoded in the weights of a set of MLPs. Then we employ a knowledge distillation strategy to convert color INR to semantic INR. In order to enhance the semantic consistency between multi-viewpoints. We design the Ray-Transformer to integrate and transfer the 3D ray-color features into ray-semantic features. Specifically, we add a CNN texture token to broadcast texture information among different locations along a ray. Finally, we combine the 3D ray-semantic features from semantic-INR with the 2D features from an additional CNN to complete the missing semantic information in novel views and get more detailed and accurate results.

Extensive experiments are conducted to verify the effectiveness of the method. We construct six sets of synthetic data based on the well-known Carla simulation platform [17]. We also construct three sets of real data from Google Maps. Our method outperforms CNN-based methods and INR-based state-of-the-art methods. The visual comparison also suggests that the proposed method can produce more accurate and visually consistent results. In addition, experiment results also show that our method has better performance against

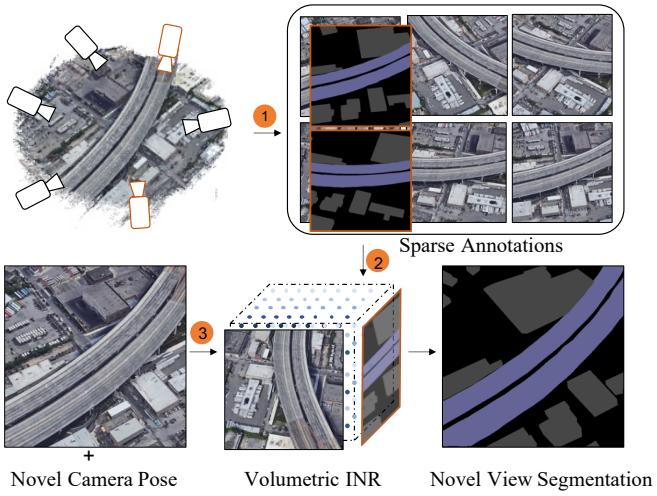

Fig. 2. In this paper, we propose an INR-based method to combine the 3D ray-semantic features and 2D CNN features for remote sensing scene semantic segmentation. ①: multi-view image capturing and sparse labeling; ②: INR construction; ③: novel view segmentation.

illumination and viewpoint changes. Our code and dataset will be made publicly available at https://qizipeng.github.io/IRT.

The contributions of this paper are summarized as follows:

- We propose a new method for multi-view remote sensing image segmentation based on the implicit neural representation, which combines the 2D CNN features with 3D ray-semantic features. Given a set of multi-view RGB images and a limited number of annotations, the proposed method effectively generates accurate segmentation results for novel views.
- We propose a density-driven and memory-friendly network architecture called Ray-Transfromer to integrate and transfer color features into semantic features; specifically, we add the CNN texture token into the Ray-Transfromer and explore a different way of introducing texture infor-

mation into the INR space.
- We construct a challenging dataset for multi-view remote sensing image segmentation, which contains both real and synthetic images. Our method reaches the state-of-the-art on the introduced dataset.

The rest of this paper is organized as follows. In Section II, we introduce the related work. In Section III, we give a detailed introduction to the proposed method. In Section IV, the experimental results are presented. Conclusions are drawn in section V.

II. Related Work

A. CNN-based Image Segmentation

In the past few years, with the development of deep learning, Convolutional Neural Networks (CNN) have become mainstream in image segmentation. These methods typically employ an encoder-decoder structure. There are mainly three groups of methods. 1) The first group adopts a Unet [4]-like architecture, where the skip connections are introduced to combine the low-level features into the decoder to keep more detailed information. Some recent methods [18–21] show the advantage of skip connections in remote sensing image segmentation tasks, *e.g.* building detection, road detection, and multi-objects change detection. 2) The second group adopts a larger respective field and a deeper architecture to extract more semantic information. The dilated convolution is employed to enlarge the respective field of the networks while maintaining a high output feature resolution [5, 22–24]. However, the stacking of multiple dilated convolutions may produce a gridding effect and thus a hybrid dilated convolution [25] is designed to alleviate this problem. 3) The third group employs a feature pyramid strategy to extract more contextual information from the images [26–29], especially for the images with multiple-scale objects. In recent years, some works [30–34] utilize the depth map which can effectively indicate the geometry information of the scene to improve the performance of single-view image segmentation and multi-view image segmentation. One of the ways to combine depth information is element-wise addition between depth features and RGB features which are extracted by a double-branch CNN backbone [32, 35]. In some work [32, 36], the fusion of two-domain (RGB and depth) information is achieved through a self-attention mechanism that re-weights features along a given dimension. In [33], authors not only consider the depth features in network architecture but use the depth map to warp the neighbor view features to a common reference view to improve the results in view consistency. The authors of [34] first render the depth map corresponding to each view and then project the 2D features to 3D space. The projected features are fused and transformed into semantic segmentation results. For multi-view image segmentation, the depth map is not only used as supplementary information but more to assist in mapping 2D features to 3D space to help the network understand the relationship between views. However, the depth maps are usually obtained by RGBD camera [37–40] or a pre-trained depth estimated network [41]. Both ways are expensive compared with only obtaining RGB images. Besides, it is difficult for a pre-trained depth estimator to have a great performance in multi-view depth prediction, which is usually trained in single-view datasets [42–44]. Note that in reality (such as Google Maps), it is difficult to obtain depth maps. Therefore, our dataset does not contain depth data. Our method utilizes an implicit neural network to extract scene geometric priors from RGB images.

B. Implicit Neural Segmentation

In the past two years, there are a variety of methods that optimize a volumetric space from a set of posed 2D images without the need for extra 3D supervision. NeRF [13] is representative of these methods. Recently, implicit neural representations have also been introduced to multi-view image segmentation tasks [38, 45–48], including indoor-scene segmentation [45], traffic-scene segmentation [46, 47] and remote sensing scene segmentation [48]. These methods can be divided into two groups. The first kind [45, 49] considers segmentation and novel view generation jointly and trains a multi-task representation where semantic features and color features share the same feature extractor but use different prediction heads. The second kind [48, 50, 51] first optimizes a color implicit neural representation of a scene and then transfers the color features into semantic features by fine-tuning or distillation. The color information can be more effectively introduced into INR space with multi-task representation. In addition to the above two groups of methods, there are also some methods [46, 52] that use extra 3D data or memory-intensive 3D convolution to improve the scene segmentation accuracy under a large number of annotations. The difference between all the above methods and our method is that we employ 3D information from the INR space and combine the CNN features with a newly designed transformer, thus achieving accurate segmentation with limited annotations.

C. Transformers

The transformer was first introduced in 2017 [53] and has been widely used in NLP tasks [54–56], which effectively solves the problem of long-range dependencies. Recently, the transformer architecture has been introduced to computer vision and remote sensing to extract global or long-range context features and shows comparable or even better performance than the CNN-based methods in various visual tasks, including image classification [57], object detection [58] and multimode tasks [2, 59, 60]. The main two components in a transformer are encoders and decoders. The transformer encoders are to explore the multi-head attention modes of input. And the transformer decoders perform the cross-attention between encoder features and additional masked input to obtain the final results. In some vision tasks, there are also some approaches that only use the transformer encoder and combine it with a CNN-based decoder. SETR [61] utilizes the transformer-based backbone and a standard CNN decoder to achieve image segmentation without decreasing the feature map resolution. Swin-transformer [62] uses a variant of ViT [57] composed of local shifting windows and a pyramid FCN decoder, which

achieves state-of-the-art performance in classification and segmentation tasks. In the remote sensing field, the transformer has many applications. The structure of the transformer is also suitable for a variety of remote sensing tasks [63–66], such as building extraction, change detection, etc. BiT [63] fuses a pair of CNN features from bi-temporal images using the transformer to improve the performance of the change-detection network. Later, SwinSUNet [64] is proposed with a pure transformer network with a Siamese U-shaped structure. Chen [67] designed the sparse token transformer to segment the buildings while greatly reducing the computational complexity in the transformer. Li [68] uses the transformer to aggregate the features of the global spatial position on multiple scales and forms a model for the interaction between instances, and finally proves that the transformer structure also has an excellent performance in remote sensing object detection. The transformer-based models usually come with larger computations and a more complex training pipeline. In our work, we design a memory-friendly transformer that works in ray space only to consider the valid 3D point features.

III. Methodology

An overview of our method is illustrated in Fig. 3. We propose an **I**mplicit **R**ay-**T**ransformer (**IRT**) for remote sensing scene semantic segmentation. The input of IRT is a group of posed images and a limited number of pixel-wise annotations. The output is the segmentation label maps of any novel views. It includes a two-stage learning process. In the first stage, we optimize a color-INR model using all posed images. The input of Φ_s is the position and view direction of sampled points along a random ray from the voxelized space of the scene. The output is the color and density attribute of each point. Then we use volume rendering to render the final pixel color of the corresponding ray. In the second stage, we distillate the point features from Φ_s and convert the color features into semantic features using a memory-friendly transformer named Ray-Tranformer. In the Ray-Transformer, we take the additional CNN features as an additional texture token to broadcast the texture information into other point features. Similarly, we render the semantic attribute of all the points along a ray and get the final ray-semantic features of the corresponding ray. In order to further complete the missing semantic information under the sparse annotations, we combine the ray-semantic features with CNN features to get the final pixel class prediction. The inference detail of IRT is shown in Algorithm 1.

Algorithm 1: Implicit Ray-Transformer for R4S

Input: $\{(\mathbf{I}_n, \mathbf{C}_n, \mathbf{L}_m)|n = 1 : N, m = 1 : M\}$ (N images(I) with camera parameters(C) and M labels (L))
Input: $\mathbf{Iteration_1}$ (iterate number in step1)
Input: $\mathbf{Iteration_2}$ (iterate number in step2)
Output: IRT
Output: $\mathbf{S}^{tgt}$ (novel view segmentation)

```
//training
  // step1: Color-INR construction
for i in 1 : Iteration_1 do
    // select a training image from I_n
    I_i, C_i = Sample(I_n, C_n)
    // select B training points in I_i with C_i
    (x, y, z, α, β)^B = SamplePoints(I_i, C_i)
    //render color results
    color^B = R(Φ_d(Φ_s(x, y, z)^B, (α, β)^B))
    //gradient descent to optimize Color-INR
end
  //step2: Seg-INR construction
for j in 1 : Iteration_2 do
    // select a training image from L_n
    L_j, C_j = Sample(L_m, C_n)
    // select B training points in I_j with C_j
    (x, y, z)^B = SamplePoints(L_j, C_j)
    //render semantic results
    //(Φ_R:Ray-Transformer, Φ_C:CNN)
    color^B = R(Φ_R(Φ_s(x, y, z)^B, Φ_C^B))
    //gradient descent to optimize Seg-INR
end
//inference
  //render segmentation result from novel view
S^tgt = IRT((x, y, z)^all)
```

A. Coordinate System Conversion

In our method, an important and fundamental operation is to sample points along a ray. Four coordinate systems and their related transformations are involved in solving the direction of the rays composed of the camera origin and the pixels in the image. The four coordinate systems include the pixel coordinate system, image coordinate system, camera coordinate system, and world coordinate system. Therefore, in this subsection, we first introduce transformations between different coordinate systems.

The pixel coordinate system represents the projection of a 3D space object on the image plane. The coordinate origin is in the upper left corner of the CCD image plane. We take (u, v) to represent the pixel coordinate. The origin of the image coordinate system is in the center of the CCD image plane, and its horizontal and vertical axes are parallel to the pixel coordinate system. We take (x, y) to represent the image coordinate. The conversion relationship between the pixel coordinate system and the image coordinate system is defined as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

where d_x and d_y respectively represent the width and height of a pixel corresponding to the photosensitive point. The camera coordinate system takes the optical center of the camera as the origin of the coordinate system. x_c, y_c axes are parallel to the x and y axes of the image coordinate system. The optical axis of the camera is set to the z_c axis and the coordinate system follows the right-hand rule.

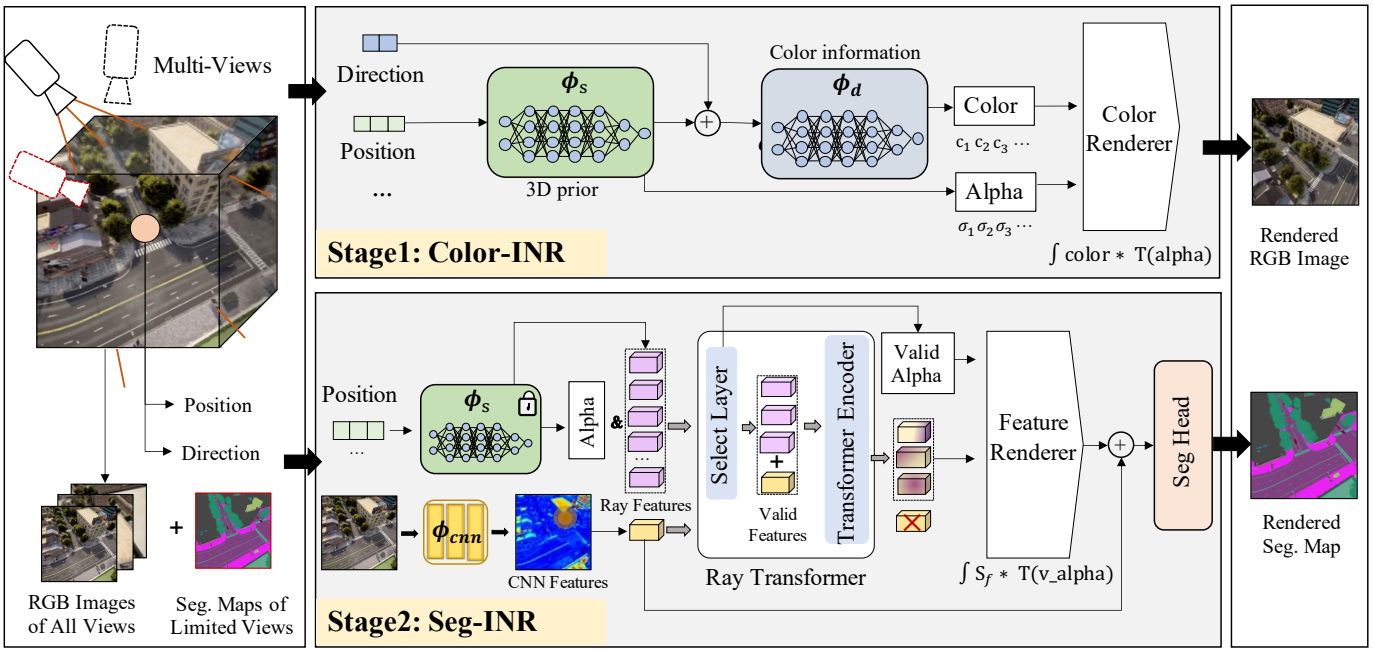

Fig. 3. The overview of the proposed model. The cameras on the left part of the figure represent different capturing positions. The black cameras represent only acquiring RGB images, and the red camera represents acquiring an RGB image and its corresponding label. The Color-INR takes all the posed images as input and optimizes the Φ_s and Φ_d to implicitly represent the scene. The 3D structure priors are encoded into the point features. The Seg-INR uses our designed Ray-Transformer to integrate and then convert the point features into ray-semantic features. In order to fully use the texture features in RGB images, we add a CNN token into the Ray-Transformer to complete information from unseen viewpoints. Finally, we can achieve novel view segmentation under sparse labels for the R4S task.

The camera coordinate system to the image coordinate system follows a perspective transformation relationship, which can be represented by using similar triangles:

$$z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & f & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = [\mathbf{K}|\mathbf{0}] \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \tag{2}$$

where f is the focal length and K is the intrinsic parameter matrix of the camera. The world coordinate system can be obtained from the camera coordinate system through rotation and translation:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{3}$$

where R is the rotation matrix and t is the translation matrix. $[R|t]$ are the extrinsic parameters of the camera. x_w, y_w, and z_w are the absolute coordinates of space objects in the world coordinate system. In our work, we use COLMAP [69] to estimate the intrinsic and extrinsic parameters of each viewpoint.

B. Color-INR

In this subsection, we introduce the details of **Color** Implicit **N**eural **R**epresentations (Color-INR), where the 3D structure information of the scene is encoded into the density attribute and the color information is encoded into the color attribute.

The Color-INR represents a scene as a neural volume field by sampling points along the rays that pass through the center of the camera and each pixel in the input image (see detail in Fig. 3). The position (x, y, z) of each sampled point and view direction (θ, β) of each ray are encoded by a positional encoding layer and then fed into a set of spatial-MLPs Φ_s and direction-MLPs Φ_d:

$$c, \sigma = \Phi_d(\Phi_s(x, y, z), \theta, \beta)) \tag{4}$$

where the output c and σ are the color attribute and density attribute respectively. The color attribute represents the (R, G, B) component and the density attribute represents the probability of light passing through this point, which can be considered as the weight of the color attribute and can be further processed to the depth attribute or the mesh results. Subsequently, we render all the density and color attributes along a ray and get the ray-color value $\hat{C}(r)$ by a discretized volume fraction:

$$\hat{C}(r) = \sum_{i=1}^{N} \exp\Big(-\sum_{j=1}^{i-1} \alpha_j \sigma_j\Big)\Big(1 - \exp\left(-\alpha_i \sigma_i\right)\Big) c_i \tag{5}$$

where α_i is the interval distance between the i point and the i+1 point. $\hat{C}(r)$ is the predicted color of the pixel on the image space corresponding to the ray r. We optimize the weights of MLPs by minimizing the sum of the square difference of all input pixel values and all rendered output:

$$\mathcal{L}_{rgb} = \sum_{r\in R} \left[\|\hat{C}(r) - C(r)\|_2^2\right] \tag{6}$$

where R is the set of all the sampling rays. After continuously optimizing the color attribute and density attribute of the sampling points in the rays, we finally get an implicit neural representation of the target scene. The geometry information of the scene is decoded into the density attribute of all sampled points. For example, we can render the depth map or the mesh result [70] only using the density attribute, as shown in Fig. 1(c). Due to the continuity of the spatial attributes in the high-dimensional space, we can also render novel view images that are not included in the training views.

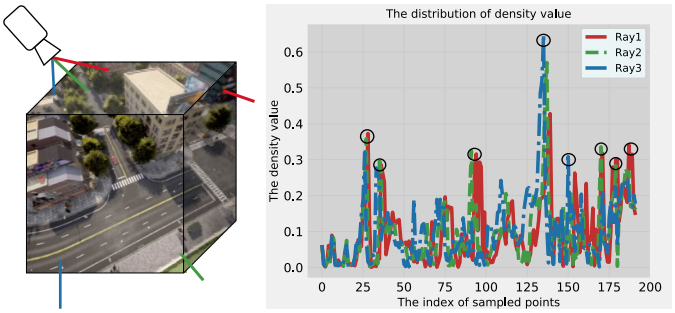

Fig. 4. The left image shows the three randomly selected rays and the right image is the density curve of the three rays. We can find that the density of most points is relatively low, which represents points in the object-free space. These points contribute very little to the result. The places with high-density values (black circles) represent the space near the surface of the object, and the places with small-density values represent the space outside or inside the object.

C. Seg-INR

In the proposed Seg-INR, we utilize the 3D structure priors, encoded in the point features, to achieve novel view segmentation. For the simplicity of the network structure design, we freeze the weights of Φ_s and then convert the 3D structure priors (the point features extracted in Φs) to the ray-semantic features. This operation also avoids repeated training of the Φ_s when adding new annotations. However, this will add difficulties to make full use of the texture in RGB images and complete the semantic information from unseen viewpoints. In addition, the color attributes change with the viewing angle, but the semantic attributes of points along a ray are ideally consistent from different viewing angles.

To address the above problems, we first eliminate the direction information (θ, β) and then design a memory-friendly Ray-Transformer to integrate the density features. Specifically, we take the CNN features as an additional token into the Ray-Transformer to broadcast the texture information. Finally, we combine the CNN features with ray-semantic features and feed the results to a Seg-Header to produce the final segmentation output.

Memory-friendly Ray-Transformer: The long-range attention in a transformer is efficient for integrating the features of points along a ray to keep semantic consistency. However, integrating all the point features along a ray will cause two drawbacks (Fig. 4): 1) The points with a low-density value in object-free space contribute less to the color-class conversion and have non-uniform semantic attributes in different rays. The object-free space refers to the space beyond objects. In this space, the sampling point is far away from the object's surface. The points in object-free space may have different semantic attributes from different view directions; 2) The computational complexity of the transformer will be greatly increased by adding more input tokens. Considering the above drawbacks, we design a memory-friendly transformer to effectively integrate and transfer the color features into semantic features.

In the memory-friendly transformer, a density-based selection layer is designed, which first selects k (k< the number of sampled points) valid points along a ray according to the density values output from the Φ_s. Then the ray features (the penultimate linear layer output from Φ_s) of the valid points are sent into a transformer encoder, which consists of multi-head self-attention (MSA) layers and MLP blocks. Specifically, we only integrate the 3D context relationship along a ray instead of in the $h \times w$ plane. At each layer, l, the query, key, and value features of each selected valid point feature F are extracted by the MLP blocks:

$$\begin{aligned} Q_l &= F \times W_l^q, \\ K_l &= F \times W_l^k, \\ V_l &= F \times W_l^v \end{aligned} \tag{7}$$

The self-attention at layer l is formulated as:

$$\mathrm{Att}(Q_l, K_l, V_l) = \mathrm{softmax}\left(\frac{Q_l K_l^T}{\sqrt{d}}\right) V_l \tag{8}$$

where d is the dimension number of features.

The self-attention mechanism globally considers all the input point features, which is suitable for keeping the consistency of semantic attributes of points along a ray. The core idea of the transformer encoder is the multi-head self-attention (MSA) which explores the multiple-mode relationship to make the model more robust. The MSA performs multiple independent self-attention heads in parallel, the outputs of the heads are concatenated and then projected into the final semantic feature of each point.

$$\begin{aligned} &\mathrm{MSA(F)} = \mathrm{concat}(\mathrm{head_1}, \mathrm{head_2}, ..., \mathrm{head_n}), \\ &\mathrm{head_j} = \mathrm{Att}(Q_l, K_l, V_l)_j \end{aligned} \tag{9}$$

The point feature integration operation in the vanilla transformer is more effective for the seen viewpoints but not efficient enough for unseen viewpoints. In order to complete the texture information from unseen viewpoints, we further design a CNN module: Φ_{cnn} as a texture extractor. We take the CNN feature point corresponding to the training ray as an additional texture token to broadcast the texture information to other tokens in the transformer.

$$s_1, s_2, \ldots, s_n = \mathrm{RT}(r_1, r_2, \ldots, r_n, c) \tag{10}$$

where $r_1, \ldots, r_n$ are the INR features and c is the CNN feature. RT is the Ray-Transformer. We name this combination of networks as Ray-Transformer with Texture (RTT) (Fig. 3).

After that, a semantic feature render takes the converted features as input and the density value from frozen Φ_s as

the weights and renders the final ray-semantic features of the training ray:

$$\hat{S}_s(r) = \sum_{i=1}^{N} \exp\Big(-\sum_{j=1}^{i-1} \alpha_j \sigma_j\Big)\Big(1 - \exp\left(-\alpha_i \sigma_i\right)\Big) s_i \quad (11)$$

where s_i is the points feature from the Ray-Transformer, α_i is the valid alpha after the selection layer. With the $\hat{S}_s(r)$, we can simply feed it into a seg-head to get the class prediction.

Semantic information completion: After the semantic information is introduced into the INR space by RTT, we explore another way to further complete the semantic information in novel viewpoints. We combine the CNN features from Φ_{cnn} with the ray-semantic features $\hat{S}_s(r)$ and we select the feature points corresponding to the training rays to align with the ray-semantic features:

$$C_s(r) = \mathrm{select_r}(\Phi_{cnn}(RGB)) \quad (12)$$

The Φ_{cnn} is supervised by the labels and generalizes to unlabeled viewpoints. The Φ_{cnn} can complete the missing information in the novel viewpoints, which is important to generate detail-rich results. Finally, we feed the fused features into a seg-head that consists of a linear prediction layer:

$$\begin{aligned} \hat{S}_c(r) &= \mathrm{concat}(\hat{S}_s(r), C_s(r)), \\ \hat{S}(r) &= \mathrm{Seg_H}(\hat{S}_c(r)) \end{aligned} \quad (13)$$

With the 3D ray-semantic features, the Φ_{cnn} easily learns robust features using a simple structure. The experiments show that the further combination can greatly improve the performance of the model.

D. Implementation Details

Network Details. In the color-INR, we adopt the two-stage point sampling strategy by following NeRF++ [14]. First, we sample 64 spatial points on a ray, and the MLP outputs the density value corresponding to each point. Given the output, we then use importance sampling to sample more spatial points near places with high-density values and fewer points near places with small-density values. Finally, 128 spatial positions are additionally sampled, resulting in 194 spatial positions. In seg-INR, we only fine-tune the point features in the second stage to accelerate training and inference. The color attribute is the three-dimensional vector and the density attribute is a one-dimensional value. The feature dimension of the penultimate linear layer in Φs is 128. In detail, the layer number of the Φs is 8 and the layer number of the Φd is 2. In the Ray-Transformer, we set the layer number of the transformer to 2 and set the number of valid points in a ray to 10. The channel number of the encoder in Ray-Transformer is 128. We use three convectional layers in the additional CNN network. The size of the convolution kernel of each layer is 3×3, and the number of channels of each layer feature is $\{3,\ 16,\ 32\}$. It is noted that the CNN network and the Ray-Transformer are trained in an end-to-end manner.

Loss function. In the Color-INR, we minimize the $\mathcal{L}_{rgb}$ to optimize the Φ_s and Φ_d. In the Seg-INR, we minimize the cross-entropy loss $\mathcal{L}_s$ to optimize the Ray-Transformer parameters. Specially, we add an extra loss $\mathcal{L}_{cnn}$ to ensure the accuracy of the CNN features. Formally, the loss function is defined as:

$$\mathcal{L}_{seg} = \mathcal{L}_s + \mathcal{L}_{cnn} \quad (14)$$

where the $\mathcal{L}_s$ and $\mathcal{L}_{cnn}$ are defined as follows:

$$\mathcal{L}_s = -\sum_{r \in R} \Big[\sum_{l=1}^{L} S^l(r) \log \hat{S}^l(r)\Big], \quad (15)$$

$$\mathcal{L}_{cnn} = -\sum_{r \in R} \Big[\sum_{l=1}^{L} S^l(r) \log {C_s}^l(r)\Big] \quad (16)$$

where L is the number of classes.

IV. Experimental Results and Analysis

A. Dataset and Metrics

In our experiment, we construct a large and challenging dataset for the multi-view remote sensing image segmentation task. In our dataset, we collect six synthetic sub-datasets using the CARLA platform and three real sub-datasets from Google Maps. We simulate the video shot around the target scene by recording the screen of Google Maps, and then take key frames of the video at equal intervals to get about 100 images from different shooting angles. For each image, we manually annotate each pixel. The results are used as the ground truth to compare the metrics of each method. For the real sub-datasets, we use the scale of Google Maps and the size of typical features to calculate that a pixel represents about 0.5 meters to 0.7 meters. CARLA is an unmanned driving simulation platform that can provide corresponding labels for the RGB images. We use CARLA's free view function to simulate drone shooting and collect RGB images and the corresponding labels to each view. Then we select about 100 images that evenly surround the target scene. The dataset contains 865 images and the size of each image is 512 $\times$ 512 pixels. Table I shows that only 2% to 6% of the images in the training set have corresponding labels. The dataset contains scenes at multiple scales, ranging from a number of buildings to an entire town. The annotations of the synthetic data are provided by CARLA, including 20 types of ground objects such as buildings, roads, zebra crossings, vegetation, and so on. The real datasets include more than 7 types of common ground objects such as buildings, roads, and vehicles. Among them, due to the automatic labeling of the CARLA platform, the labels in the synthetic datasets are more refined (The category details can be viewed on our project homepage).

Our method samples spatial points in the space corresponding to all views in the Color-INR stage, and uses RGB images to optimize the density attribute and color attribute of each point. In the Seg-INR stage, we sample spatial points in the space belonging to a small number of selected views and optimize the Ray-Transformer with labels corresponding to these views. Based on the above method, our datasets pose three challenges to the multi-view image segmentation task. The first challenge is the limited annotations, where the labeled images only occupy 2% - 6% of the whole training set. The second challenge is the category imbalance problem. The third

TABLE I
DETAILS OF THE SYNTHETIC AND REAL DATASET.

	#views	labeling ratio	#classes
sys #1	100	3%	20
sys #2	100	4%	18
sys #3	100	5%	20
sys #4	80	6%	19
sys #5	85	6%	19
sys #6	100	5%	18
real #1	100	2%	5
real #2	100	2%	3
real #3	100	2%	3

challenge is that different ground objects may have similar textures and contain some densely located ground objects, which are difficult to distinguish, as shown in Fig. 5. We can find that building roofs and roads have similar textures, and there are also some densely parked cars.

Fig. 5. The first row are the data samples in sys #1; The second row are the data samples in sys #2; The last row are the data samples in real #1.

In our experiment, we use mean Intersection over Union (MIoU), the most commonly used segmentation metric, to compare the performance of different methods:

$$MIoU = \frac{TP}{FP + FN + TP} \tag{17}$$

where the TP is the true positive, FP is the false positive, FN is the false negative.

B. Comparison Methods

In order to verify the effectiveness of our method, we choose the following methods for comparative experiments. All comparative methods are retrained on the proposed datasets to ensure fairness. Besides, we also compare different strategies to introduce texture information into the INR space. The baseline strategy [45] performs both color reconstruction and semantic segmentation. It utilizes two MLP heads to process the spatial features and simultaneously predict RGB values and semantic values. In final, the baseline strategy uses a uniform renderer function to render color results and semantic segmentation. Another two-stage strategy is to reconstruct color information first and then convert it into segmentation [48]. We also compare three different approaches to using a transformer to integrate point features based on the above two-stage strategy. The first approach is to use a transformer to simply integrate long-range point features along a training ray to obtain semantic features; the second approach is to design a spatial CNN token and broadcast texture information from an input RGB image into other tokens; the final way is to further concatenate the CNN features with integrated features by the transformer to enhance the detail of the texture information. In addition to the above variants, we also conduct a horizontal comparison with methods published in recent years. The comparison methods are as follows:

1) SegNet [71]: a CNN-based semantic segmentation method that uses an encoder-decoder architecture;
2) Unet [4]: a CNN-based semantic segmentation method that adopts the skip-connection to keep the detailed information in encoder layers;
3) DANet [6]: a CNN-based semantic segmentation method that proposes a dual attention mechanism to adaptively integrate local features and global dependencies;
4) DeepLabv3 [5]: a CNN-based semantic segmentation method that uses the dilated convolutions to effectively enlarge the receptive field;
5) SETR [61]: a transformer-based semantic segmentation method;
6) Sem-NeRF [45]: the first NeRF-based model for indoor image semantic segmentation;
7) Color-NeRF [48]: a NeRF-based method for semantic segmentation that adds an additional color-radiance network to fuse the pixel-level color information for improving the NeRF-based segmentation;
8) IRT (B): Our baseline method which transforms the color features into semantic features only by fine-tuning;
9) IRT (RT_2): A variant of our method that integrates the point features along a ray based on the baseline by a two-layer transformer;
10) IRT (RT_6): A variant of our method that has a similar configuration with IRT (RT_2) except the number of transformer layers is set to 6;
11) IRT (RTT): A variant of our method where we take the CNN features as semantic guidance based on IRT (RT_2);
12) IRT (RTC): A variant of our method where we combine the ray features and CNN features produced by the Ray-Transformer based on IRT (RT_2);
13) IRT (RTTC): A variant of our method where we combine the ray features and CNN features produced by the RTT.

C. Overall Comparison Results

Qualitative and Quantitative Results In Fig. 6, we show the qualitative results of CNN-based and INR-based methods. The results show that the INR-based methods get better results than the CNN-based methods. Although Unet has the best performance among all CNN-based methods, it still fails in

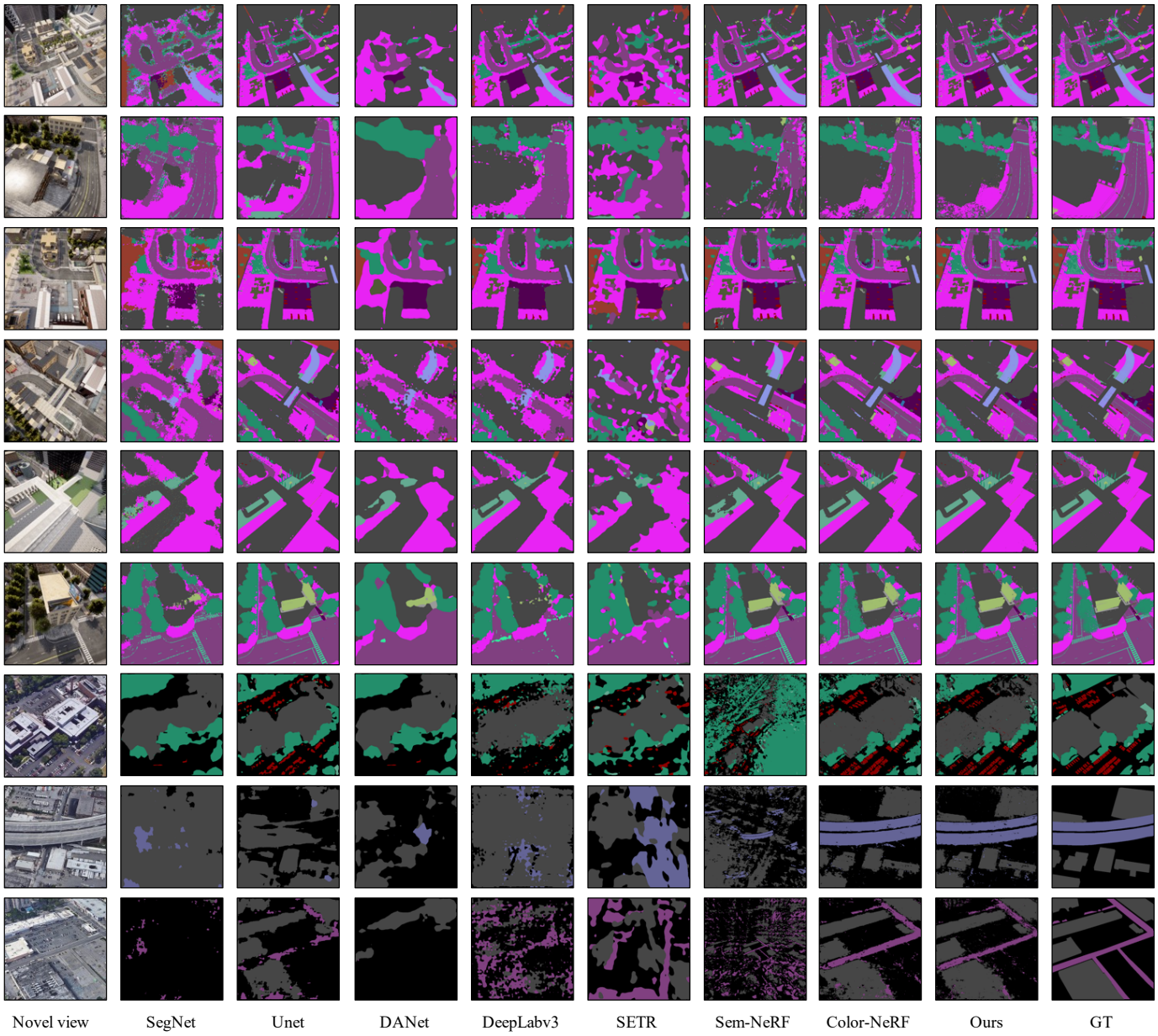

Fig. 6. In the CNN-based methods, Unet has the best performance but it still fails in real scenes. Compared to other methods, our method generates more accurate and complete results, such as zebra crossings, small trees, and traffic lights. The advantage is the introduction of texture information from the Ray-Transformer. At the same time, our method achieves more robustness in the multi-scale object. This is due to the continuity of spatial attributes in the high dimension. From top to bottom, the images of each row are from scenes sys #1, sys #2, sys #3, sys #4, sys #5, sys #6, real #1, real #2, real #3

the datasets of objects with similar textures. The key reason is that CNN-based methods only emphasize two-dimensional information. This leads to these methods requiring a large number of annotations to extract discriminative texture features to distinguish similar ground objects. For example, in real #2, the background and the bridges have similar textures, and the Unet cannot distinguish the two types of ground features well. In real #3, the roads and the background are almost indistinguishable. If only using the 2D texture features, the Unet cannot generate the complete path. On the contrary, in real #1, the textures of buildings, backgrounds, and other subclasses vary greatly, so the Unet performs better on this dataset. The above results also verify the correctness of our analysis. Our method with the implicit 3D information from Color-INR is more powerful to distinguish similar textures for various scenes. The key advantage is that we take the 3D structure into consideration to help the model distinguish the objects in different depths. Compared to Sem-NeRF, our method can generate more detailed and clearer segmentation of small-scale objects, such as zebra crossing, small trees, and traffic-lights. Our method also successfully achieves dense object segmentation in real #1. At the same time, our method can generate complete and accurate big-scale object segmentation, such as the building in sys #5. The adaptability of

TABLE II
MIoU METRIC OF DIFFERENT METHODS ON EACH SUB-DATASET.

Methods	sys #1	sys #2	sys #3	sys #4	sys #5	sys #6	real #1	real #2	real #3	AVG
SegNet	11.79	13.21	10.81	26.71	8.21	18.77	8.67	27.88	18.84	16.10
Unet	23.92	31.73	42.26	41.79	26.63	38.72	64.94	49.95	68.33	43.14
DANet	9.73	13.91	16.62	26.78	8.53	15.79	35.71	49.13	34.30	23.39
Deeplab	18.89	16.64	19.96	30.42	11.90	20.52	39.35	41.37	52.37	27.94
SETR	10.34	10.71	12.45	21.13	8.90	13.97	36.26	31.04	26.07	18.99
Sem-NeRF	55.73	34.81	49.82	57.24	41.44	41.85	11.64	18.78	19.76	36.79
Color-NeRF	**58.03**	38.46	50.86	59.14	**43.77**	43.53	62.04	85.85	69.92	56.84
IRT(B)	53.12	33.95	44.69	55.30	39.08	41.20	60.65	83.40	69.37	53.42
IRT(RT_2)	53.70	34.97	44.97	55.47	39.09	41.50	60.96	83.58	70.98	53.91
IRT(RT_6)	52.70	34.97	44.60	55.65	39.57	42.31	60.51	83.57	71.37	53.92
IRT(RTT)	54.39	41.65	51.79	57.98	41.28	46.99	65.61	85.98	**71.71**	57.49
IRT(RTC)	57.61	42.19	49.43	**60.33**	42.43	45.15	65.51	**86.06**	62.06	56.75
IRT(RTTC)	57.86	**43.23**	**53.47**	59.98	43.73	**48.59**	**65.84**	84.31	71.02	**58.67**

our method at multiple scales is due to the continuity of high-dimensional spatial properties, which do not depend on the resolution. It is noticed that Sem-INR fails in real sub-datasets. This is due to the poor ability of the vanilla NeRF-based method to reconstruct the scenes with larger differences between foreground and background.

From Table II, we can find that the INR-based methods get a higher MIoU score than CNN-based methods in all sub-datasets. Compared to Unet which has the best performance in CNN-based methods, the proposed IRT outperforms the Unet by 15.5% on the average MIoU. Compared to the Sem-NeRF, our method outperforms it by 21.87% in average MIoU. Our results also outperform the Color-NeRF by 1.81% in average MIoU. Our method shows greater superiority on synthetic data, suggesting that IRT is more robust to the cases with more classes and detailed annotations. Further comparisons along synthetic sub-datasets show that with the introduction of texture information, our model further improves on generating detailed results for high-resolution images.

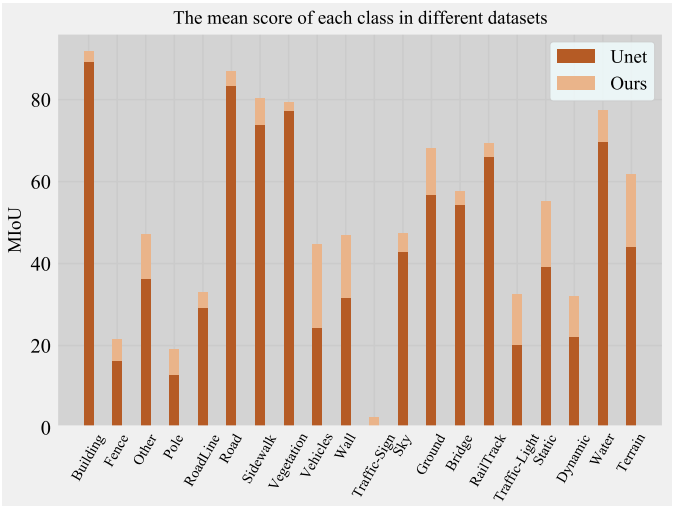

Fig. 7. The avg MIoU of each class in sub-datasets. The result shows that the proposed method is more friendly to the classes with a small number of annotations.

In Fig. 7 and Fig. 8, we show the statistical results at the category level and dataset level. Fig. 7 shows the average MIoU across six synthetic sub-datasets for each category. Taking the building category as an example, the statistical method is $\frac{1}{6}\sum_{i=1}^{6} MIoU_i$. $MIoU_i$ represents the MIoU metric of the building category in i_{th} synthetic sub-dataset. The labels of the synthetic sub-datasets are provided by the CARLA platform with unified semantics but our manually annotated real datasets do not have unified semantics. In addition, compared with real datasets, synthetic datasets have more obvious differences in the number distribution of different categories and contain more subcategories. Based on the above, we only perform statistics on synthetic sub-datasets. With Fig. 7, we find that the category with a large number of samples usually has a higher value of MIoU. We further find that our method has more advantages in classifying categories with a small number of samples and even with only a few pixel labels (*e.g.*, Traffic-Sign category). In addition, we also counted the distribution of the different methods' MIoU metric in nine sub-datasets and make a box plot for display. Fig. 8 shows that our method generally performs better than other methods. Comparing Color-NeRF which is close to our value, we can find that our method has a higher lower line than it, and the index values of different sub-datasets are more concentrated, which means our method is more stable in different scenarios.

Semantic Consistency. We compare the consistency of the segmentation results of different methods, including Unet, Sem-NeRF, Color-NeRF, and ours. In Fig. 9, we can see that Unet has low accuracy in classifying buildings from the first few viewpoints, and since it processes images from each viewpoint separately, it does not maintain view consistency. On the contrary, our methods work in the 3D INR space and mainly take the location of spatial points as the input, therefore, presenting better view consistency. Color-NeRF will misclassify buildings (gray labels) as roads (purple labels), and the results are less consistent in terms of perspective. Since our method incorporates CNN features, it can achieve more accurate results compared to Color-NeRF which only extracts pixel color information. In addition, we can also find that our method has a better representation of the structure of the object. For example, the structure of the building is more complete. The reason is that we introduce a transformer struc-

TABLE III
QUANTITATIVE COMPARISON OF USING DIFFERENT VALID POINTS IN THE RAY-TRANSFORMER.

number of points	sys #1	sys #2	sys #3	sys #4	sys #5	sys #6	real #1	real #2	real #3
10 points	57.86	43.23	53.47	59.98	43.73	48.59	65.85	84.31	71.02
20 points	57.97	43.28	53.22	60.18	43.99	48.67	65.00	84.46	71.81

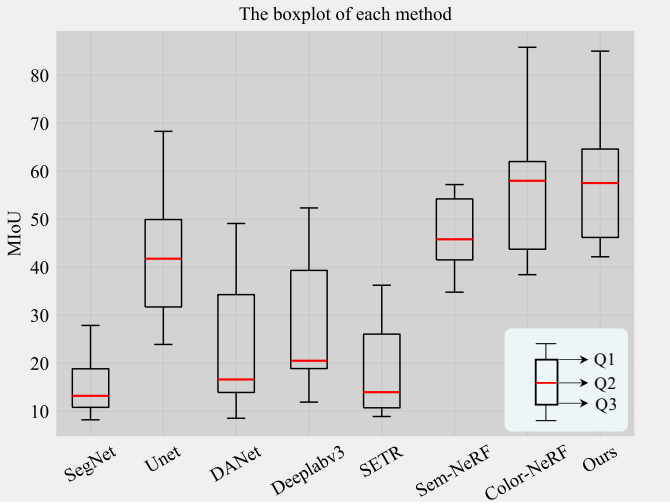

Fig. 8. The boxplot results of all methods. The Q1, Q2, and Q3 values of our method are higher. Q1: first quartile value of the result; Q2: the median value of the result; Q3: the last quartile value of the results.

ture to enhance the semantic consistency of spatial points in a ray, and we design a selection layer to remove the influence of object-free points. Besides, since our method incorporates CNN features, it can achieve more accurate results compared to Color-NeRF which only extracts pixel color information.

D. Ablation Study

In this part, we conduct ablation experiments, mainly to verify 1) the necessity of the CNN token, 2) the effectiveness of the number of valid points in the Ray-Transformer, and 3) the necessity of the Color-INR. Our evaluation results are shown in the lower part of Table II and Table III. The visual results are shown in Fig. 10.

Ray-Transformer and CNN token. In Table II, the IRT (RT_2) and IRT (RT_6) represent the vallina Ray-Transformer with different encoder layers. Compared to the baseline model, we can see that the vanilla Ray-Transformer only brings minor improvement and the deeper transformer structure can not further improve the performance. The reason is that the vanilla Ray-Transformer, which mainly integrates the long-range relationship of input tokens in ray space generates more view consistency and detailed results but can not further extract texture information to improve the performance greatly.

Although the vanilla transformer can not introduce extra information, its structure is suitable for fusing different information in different feature spaces, such as INR feature space and CNN feature space. The accuracy boost of IRT (RTT) suggests that the Ray-Transformer taking CNN tokens as input is important for completing the missing information of novel views under the sparse annotations. In addition, we also try to introduce the texture features by combining the CNN feature with the vanilla Ray-Transformer features (RTC). The result is still improved, but not as much as the IRT (RTT). This demonstrates the effectiveness of the transformer architecture.

With the combination of the CNN and Transformer, the performance has been further improved and reached the SOTA. In detail, we find that adding texture features is more effective for the small-scale scenes, *e.g.* sys #2 and sys #3. We also compare the difference between only adding the CNN token into Ray-Transformer (IRT(RTT)) with combining the CNN feature with ray-semantic features (IRT(RTC)). The results show that broadcasting texture information in ray space is more effective for small-scale scenes. The CNN features with a larger receptive field are more effective in large-scale scenes, *e.g.*, sys #1.

The number of sampling points. we explored the influence of the number of sampling points input in the Ray-Transformer on the final segmentation result. As shown in Table III, we tested two different settings: there are 10 or 20 sampling points on a ray. we found out that increasing the number of points from 10 to 20 improved the results, but very slightly. This means that even if the number of spatial points sent to the Ray-Transformer increases, the points belonging to the object-free space among these points will have a very limited contribution to the result, and will instead increase computing resources and time exponentially. This further verifies the effectiveness of the selection layer we added to the transformer structure.

The necessity of Color-INR. Fig. 10 shows the results with and without Color-INR. Without extracting 3D priors with Color-INR, we directly train Seg-INR using only a small number of labels. From the results, it can be found that a rough segmentation result can be obtained without Color-INR. For example, most buildings and roads can be identified. But it is difficult to identify some ground objects without obvious distinguishing features. Furthermore, without 3D priors, it is difficult for the network to obtain spatially continuous features. For this reason, there are many holes and discontinuous semantic labels in the result. The inaccuracy of spatial features will further lead to the inability to distinguish similar ground objects well. This result can also further validate that 3D information is crucial for scene segmentation tasks with only a small amount of supervised information.

E. Robust again Illumination and View Changes

In this part, we compare our method against CNN-based methods on the robustness against illumination and view changes.

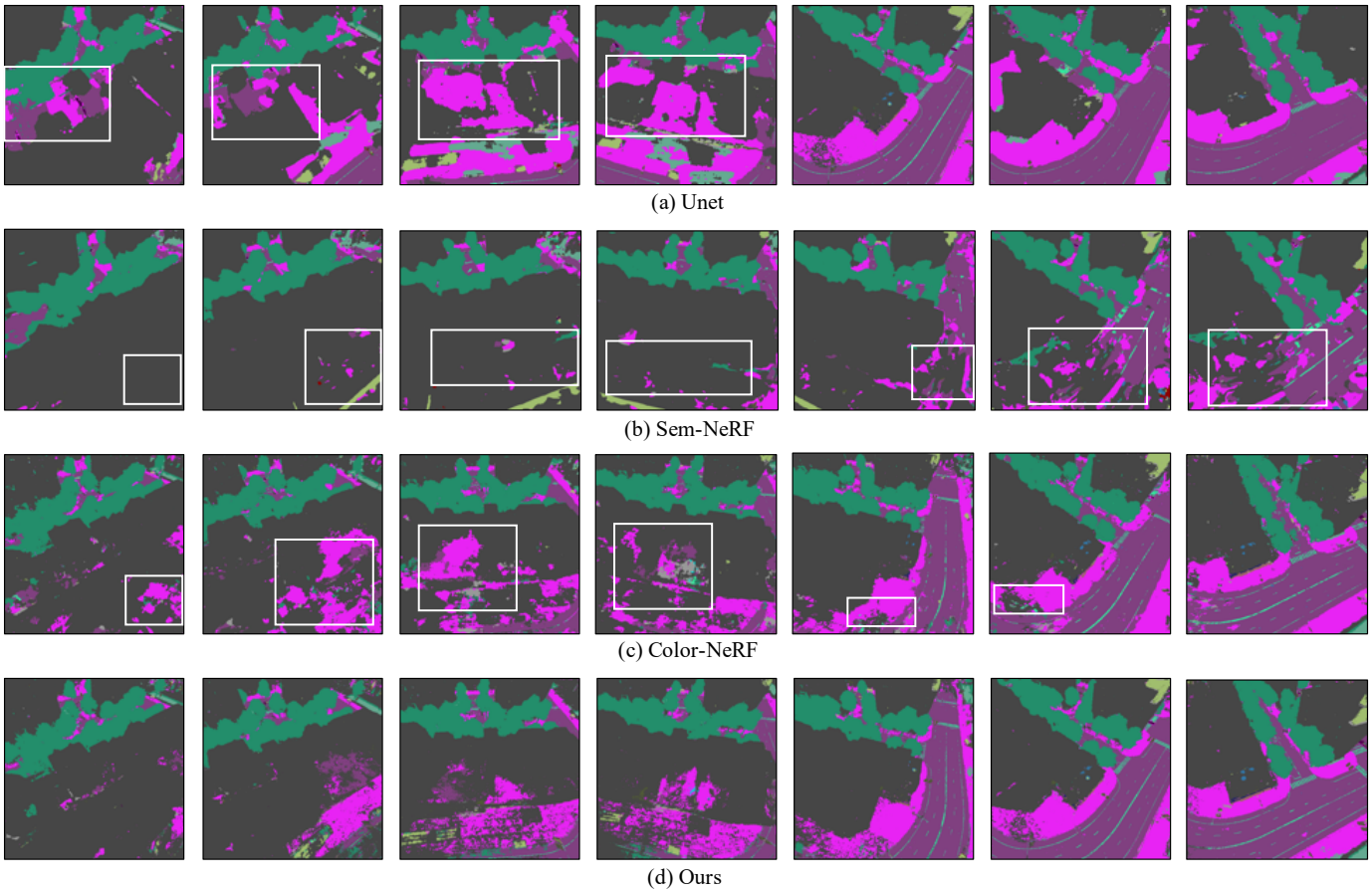

Fig. 9. The semantic-consistent results from sys #2. The results show that IRT can generate results with more accuracy and view consistency. The labels in white boxes are not accurate or view-consistent.

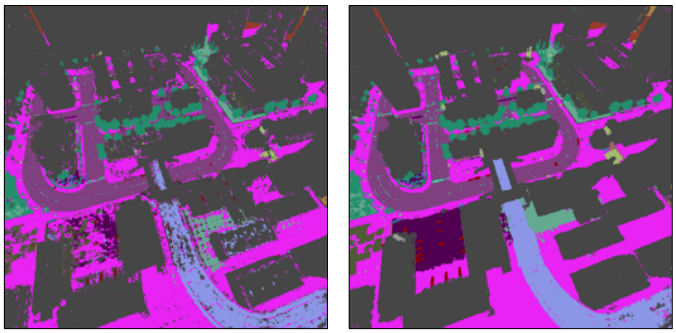

Fig. 10. The left image shows the result without Color-INR and the right image shows the result with Color-INR. Comparing the two results, we can find that the Color-INR which encodes the 3D structure of the scene can effectively keep the spatial-semantic continuity of the result.

Robustness against illumination changes. We randomly change the intensity of the input images to simulate lighting changes and test directly on the non-retrained CNN-based models and our model. As shown in Table IV, the CNN-based methods show weak adaptability to illumination changes. In contrast, our method maintains the highest accuracy and the smallest precision fluctuation among all methods. The visualized result (Fig. 11) also suggests the above conclusion. The robustness of our method mainly comes from the introduced 3D structural information, which does not change with illumination.

Robustness against view changes. We test the robustness of our method on the novel views that were completely uncovered by the training dataset. We use the viewpoints of sys #2 to test the model trained in sys #1. Our model discretizes the target scene by sampling spatial points on each ray in the camera-to-ground direction. The model encodes the location of each spatial point and generates corresponding density and color attribute values. In the Seg-INR stage, spatial point density features are used as 3D priors to generate corresponding semantic features. Different camera positions correspond to different sampling points. In our dataset, 6 synthetic sub-datasets cover shooting data at different heights and locations. From Figure 6, it can be found that the heights taken in sys #1 and sys #2 are very different. We find that although the model is trained on a high-altitude view (from sys #1), it can also be directly tested on a never-trained low-altitude view (from sys #2) with good results. As shown in Fig. 12, although our model is trained on high-altitude scenes, it can still adapt to low-altitude viewpoints and generate accurate results. At the same time, it also means that our method has greater potential

TABLE IV
MIOU METRIC OF DIFFERENT METHODS UNDER DARK ENVIRONMENT. MIOU/PERFORMANCE DROP(%)

Sub-Scenes	SegNet		Unet		DANet		Deeplab		SETR		Ours	
sys #1	1.86	-84.22%	4.46	-81.35%	3.88	-60.12%	7.41	-60.70%	3.12	-69.82%	**46.17**	**-20.20%**
sys #2	3.51	-73.42%	6.82	-78.50%	5.32	-61.75%	5.17	-68.93%	4.57	-57.32%	**34.34**	**-20.56%**
sys #3	2.11	-80.48%	4.82	-88.59%	5.21	-68.65%	6.37	-68.08%	4.57	-63.29%	**39.23**	**-26.63%**
sys #4	2.42	-90.93%	7.12	-82.96%	8.94	-66.61%	10.23	-66.37%	4.58	-78.32%	**49.53**	**-17.42%**
sys #5	3.46	-57.85%	4.54	-82.95%	3.17	-62.83%	4.87	-59.07%	3.78	-57.52%	**29.70**	**-32.08%**
sys #6	5.40	-71.23%	4.86	-87.44%	5.87	-62.82%	7.35	-64.18%	6.63	-52.54%	**35.07**	**-27.82%**
real #1	5.01	-42.21%	8.29	-87.23%	15.56	-56.42%	15.76	-59.94%	7.58	-79.09%	**58.56**	**-11.05%**
real #2	15.20	-45.48%	19.88	-60.20%	20.75	-57.76%	28.10	-32.07%	19.07	-38.56%	**82.94**	**-1.62%**
real #3	16.75	-11.09%	15.83	-76.83%	19.18	-44.08%	29.60	-43.47%	9.08	-65.17%	**67.90**	**-4.39%**
AVG	6.19	-61.88%	8.51	-80.67%	9.76	-60.12%	12.76	58.09%	7.01	-62.40%	**49.27**	**-17.97%**

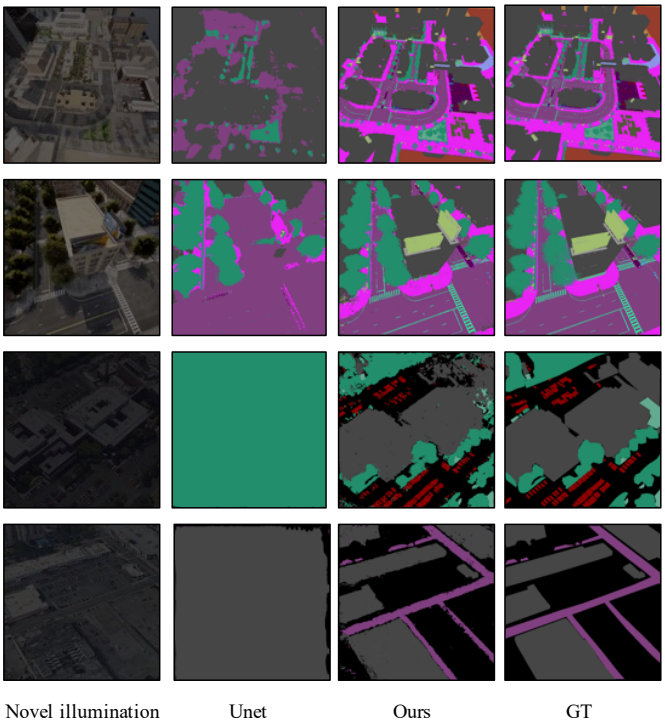

Fig. 11. We randomly change the intensity of the input images to simulate lighting conditions in a dark environment and test them with non-retrained models. The results show that our method significantly outperforms the CNN-based method.

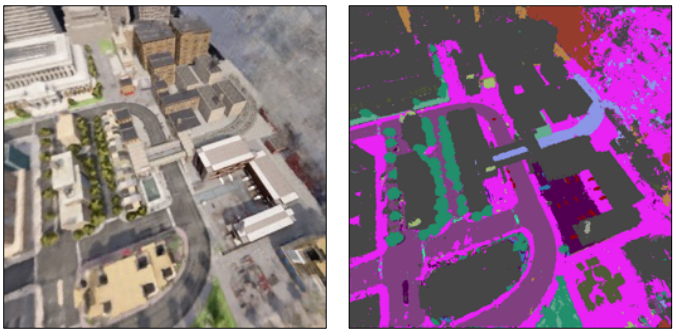

Fig. 12. Our model is trained on sys #1 views and directly tested on view from sys #2 . The left part shows the corresponding RGB image that we tested directly on the untrained view, and the right part shows the result of our segmentation of the left image.

in handling images of different resolutions caused by shooting at different heights.

V. CONCLUSION

In this paper, we consider multi-view remote sensing image segmentation under sparse annotations and propose a new method based on implicit neural representations and transformers. We optimize the implicit volume representation of the 3D scene by fitting the posed RGB images into a neural network. Then a Ray-Transformer network combines the CNN features with the 3D volume representation to complete the missing information of the unknown views. To achieve this, we also introduce a challenging dataset for the R4S task. Extensive experimental results verify the effectiveness of our proposed method. The results demonstrate that our method outperforms other CNN-based methods in terms of both accuracy and robustness. We also compare different strategies to add texture information into INR feature space and show the effectiveness of the transformer structure for this task. Finally, our empirical results also indicate the robustness of our method against illumination and viewpoint changes in the scene.

REFERENCES

[1] W. Li, Z. Zou, and Z. Shi, "Deep matting for cloud detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8490–8502, 2020.

[2] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[3] H. A. Al-Najjar, B. Kalantar, B. Pradhan, V. Saeidi, A. A. Halin, N. Ueda, and S. Mansor, "Land cover classification from fused dsm and uav images using convolutional neural networks," *Remote Sensing*, vol. 11, no. 12, p. 1461, 2019.

[4] 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.

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[6] 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.

[7] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, "Deep mesh reconstruction from single rgb images via topology modification networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9964–9973.

[8] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik, "Learning category-specific mesh reconstruction from image collections," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 371–386.

[9] C. Poullis and S. You, "Automatic reconstruction of cities from remote sensor data," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 2775–2782.

[10] Y. Ye, S. Tulsiani, and A. Gupta, "Shelf-supervised mesh prediction in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8843–8852.

[11] Z. Hu, X. Bai, J. Shang, R. Zhang, J. Dong, X. Wang, G. Sun, H. Fu, and C.-L. Tai, "Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 488–15 498.

[12] Y. Song, F. He, Y. Duan, T. Si, and J. Bai, "Lslpct: An enhanced local semantic learning transformer for 3-d point cloud analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[13] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

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

[15] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[16] Y. Wu, Z. Zou, and Z. Shi, "Remote sensing novel view synthesis with implicit multiplane representations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[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] L. Yuan, Y. Li, Y. Si, J. Ren, Y. Yang, Y. Gong, Y. Xia, Z. Tong, and L. Tong, "Multi-objects change detection based on res-unet," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 4364–4367.

[19] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[20] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[21] A. Ghosh, M. Ehrlich, S. Shah, L. S. Davis, and R. Chellappa, "Stacked u-nets for ground material segmentation in remote sensing imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 257–261.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[23] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 1442–1450.

[24] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7503–7520, 2019.

[25] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*. Ieee, 2018, pp. 1451–1460.

[26] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[27] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sensing*, vol. 9, no. 5, p. 446, 2017.

[28] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, 2020.

[29] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2245–2255, 2020.

[30] R. Schön, K. Ludwig, and R. Lienhart, "Impact of pseudo depth on open world object segmentation with minimal user guidance," *arXiv preprint arXiv:2304.05716*, 2023.

[31] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for rgb-d salient object detection," *IEEE Transactions on Image Processing*,

vol. 30, pp. 1949–1961, 2021.

[32] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1440–1444.

[33] L. Ma, J. Stückler, C. Kerl, and D. Cremers, "Multi-view deep learning for consistent semantic mapping with rgb-d cameras," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 598–605.

[34] A. Kundu, X. Yin, A. Fathi, D. Ross, B. Brewington, T. Funkhouser, and C. Pantofaru, "Virtual multi-view fusion for 3d semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 518–535.

[35] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Citeseer, 2010, pp. 708–721.

[36] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Springer, 2020, pp. 561–577.

[37] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7283–7290.

[38] Z. Murez, T. v. As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *European conference on computer vision*. Springer, 2020, pp. 414–431.

[39] 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.

[40] Z. Qi, Z. Zou, H. Chen, and Z. Shi, "3d reconstruction of remote sensing mountain areas with tsdf-based neural networks," *Remote Sensing*, vol. 14, no. 17, p. 4333, 2022.

[41] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen, "Learning to recover 3d scene shape from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 204–213.

[42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images." *ECCV (5)*, vol. 7576, pp. 746–760, 2012.

[43] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[44] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1625–1632.

[45] 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.

[46] X. Fu, S. Zhang, T. Chen, Y. Lu, L. Zhu, X. Zhou, A. Geiger, and Y. Liao, "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation," *arXiv preprint arXiv:2203.15224*, 2022.

[47] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. J. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser, "Panoptic neural fields: A semantic object-aware neural scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 871–12 881.

[48] Z. Qi, Z. Zou, H. Chen, and Z. Shi, "Remote sensing image segmentation based on implicit 3d scene representation," *IEEE Geoscience and Remote Sensing Letters*, 2022.

[49] W.-C. Tseng, H.-J. Liao, L. Yen-Chen, and M. Sun, "Clanerf: Category-level articulated neural radiance field," *arXiv preprint arXiv:2202.00181*, 2022.

[50] C. Eteke, J. Zhang, and E. Steinbach, "Semantic-srf: Sparse multi-view indoor semantic segmentation with stereo neural radiance fields," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[51] S. Vora, N. Radwan, K. Greff, H. Meyer, K. Genova, M. S. Sajjadi, E. Pot, A. Tagliasacchi, and D. Duckworth, "Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes," *arXiv preprint arXiv:2111.13260*, 2021.

[52] V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi, "Neural feature fusion fields: 3d distillation of self-supervised 2d image representations," *arXiv preprint arXiv:2209.03494*, 2022.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[55] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[56] W. Fedus, B. Zoph, and N. Shazeer, "Switch transform-

ers: Scaling to trillion parameter models with simple and efficient sparsity."
[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[58] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
[59] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[60] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1439–1449.
[61] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
[62] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
[63] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
[64] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
[65] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[66] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.
[67] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sensing*, vol. 13, no. 21, p. 4441, 2021.
[68] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer cnn for remote-sensing-image object detection," *Remote Sensing*, vol. 14, no. 4, p. 984, 2022.
[69] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.
[70] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao, "Nerf-editing: geometry editing of neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 353–18 364.
[71] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

Zipeng Qi received his B.S degree from Hebei University of Technology in 2018. He is currently pursuing his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include image processing, deep learning, and pattern recognition.

Hao Chen received his B.S. degree from the Image Processing Center School of Astronautics, Beihang University in 2017. He is currently pursuing his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University.His research interests include machine learning, deep learning and semantic segmentation.

Chenyang Liu received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently working towards the M.S. degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include machine learning, computer vision and multimodal learning.

Zhenwei Shi (M'13) received his Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005. He was a Postdoctoral Researcher in the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar in the Department of Electrical Engineering and Computer Science, at Northwestern University, U.S.A., from 2013 to 2014. He is currently a professor and the dean of the Image Processing Center, School of Astronautics, Beihang University. His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Associate Editor for the *Infrared Physics and Technology*. He has authored or co-authored over 100 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Neural Networks, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters and the IEEE Conference on Computer Vision and Pattern Recognition. His personal website is http://levir.buaa.edu.cn/.

Zhengxia Zou received his B.S. degree and his Ph.D. degree from the Image Processing Center, School of Astronautics, Beihang University in 2013 and 2018, respectively. He is currently a Professor at the School of Astronautics, Beihang University. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include remote sensing image processing, computer vision, and related problems in autonomous driving and games. He has published more than 40 peer-reviewed papers in top-tier journals and conferences, including Nature Communications, Proceedings of the IEEE, IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, and IEEE / CVF Computer Vision and Pattern Recognition. Zhengxia Zou was selected as "World's Top 2% Scientists" by Stanford University in 2022. His personal website is https://zhengxiazou.github.io/.