文章编号：1001-9014(XXXX)XX-0001-13

DOI:10.11972/j.issn.1001-9014.XXXX.XX.001

Infrared-NeRF: a low resolution thermal infrared light field 3D reconstruction method based on NeRF

Huang Yi-Fan[1], Wang Rui[2], Deng Li-Ming[2], LI Jia-Jia[2], Li Xi-Cai[2*]

(1. School of Computer Science, Nanjing University, Nanjing 210023, China;
2. School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China)

Abstract: This article proposes a three-dimensional light field reconstruction method based on neural radiation field (NeRF) called Infrared NeRF for low resolution thermal infrared scenes. Based on the characteristics of low resolution thermal infrared imaging, various optimizations have been carried out to improve the speed and accuracy of thermal infrared 3D reconstruction. Firstly, inspired by Boltzmann´s law of thermal radiation, distance was incorporated into the NeRF model for the first time, resulting in a nonlinear propagation of a single ray and a more accurate description of the physical property that infrared radiation intensity decreases with increasing distance. Secondly, in terms of improving inference speed, based on the phenomenon of high and low frequency distribution of foreground and background in infrared images, a multi ray non-uniform light synthesis strategy is proposed to make the model pay more attention to foreground objects in the scene, reduce the distribution of light in the background, and significantly reduce training time without reducing accuracy. In addition, compared to visible light scenes, infrared images only have a single channel, so fewer network parameters are required. Experiments using the same training data and data filtering method showed that compared to the original NeRF, the improved network achieved an average improvement of 13.8% and 4.62% in PSNR and SSIM, respectively, while an average decrease of 46% in LPIPS. And thanks to the optimization of network layers and data filtering methods, training only takes about 25% of the original method´s time to achieve convergence. Finally, for scenes with weak backgrounds, this article improves the inference speed of the model by 4-6 times compared to the original NeRF by limiting the query interval of the model.

Key words: neural radiation field, 3D reconstruction, thermal infrared NeRF, foreground segmentation, low resolution

PACS:

Infrared-NeRF:一种基于NeRF的低分辨率热红外光场3D重建方法

黄一凡[1], 汪 锐[2], 邓力鸣[2], 李佳家[2], 李希才[2*]

(1. 南京大学 计算机科学与技术学院,江苏 南京 210023;
2. 南京大学 电子科学与工程学院,江苏 南京 210023)

摘要:本文针对低分辨率热红外场景提出了一种基于神经辐射场(NeRF)的三维光场重建方法Infrared-NeRF。结合低分辨率热红外成像特点,面向热红外三维重建的速度和精度提升问题开展了多方面优化。首先,受玻尔兹曼热辐射定律启发,首次将距离这一因素纳入NeRF模型中,使得单光线的传播呈非线性变化,更为准确地描述了红外辐射强度随距离增大而递减的物理性质。其次,在推理速度提升方面,基于红外图像中前景和背景呈高低频分布现象,提出了多光线非均匀光线合成策略,使得模型更关注场景中的前景物体,减少背景中光线的分布,在不降低精度的情况下大幅减少训练时间。此外,相较于可见光场景红外图像仅有单通道,因此只需要较少的网络参数。经过同一训练数据和数据筛选方法进行实验表明,相较原始NeRF,改进后的网络PSNR和SSIM平均提升了13.8%和4.62%,LPIPS平均降低了46%。并且得益于网络层数的优化和数据甄别方法,训练仅需耗费约原方法25%的时间即可达到收敛。最后,针对背景暗弱的场景,本文通过限制模型查询区间的方法使得模型的推理速度相较于原始NeRF提升了4-6倍。

关 键 词:神经辐射场;三维重建;热红外NeRF;前景分割;低分辨率

Foundation items: This research was support by The Fundamental Research Funds for the Central Universities (Integrated Innovation Category 2024300443) and the National Natural Science Foundation of China (NSFC) Young Scientists Fund (Grants No 62405131).

中图分类号:TP18

Introduction

Infrared 3D reconstruction is a supplement to visible light 3D reconstruction, which can not only obtain more information that is not visible at the visible light level, but also obtain depth maps of infrared images to understand the positional relationships between objects in the scene. This is crucial for understanding and analyzing scene structures, therefore infrared 3D reconstruction has important applications in many fields. In the field of autonomous driving, infrared 3D reconstruction technology can help vehicles and robots obtain more comprehensive environmental information[1], including the position, shape, and structure of objects, thereby achieving more accurate navigation and obstacle avoidance. In the field of medicine, infrared 3D reconstruction technology can display the patient's physical condition in detail, which helps doctors quickly and accurately determine the condition of lesions. Especially in the diagnosis and treatment of tumors, infrared thermography has significant advantages[2]. In terms of security monitoring, infrared 3D reconstruction technology can display the temperature distribution field of objects, transforming the temperature distribution in the non visible light band into a thermal map of the target surface temperature distribution that can be recognized and visualized by the human eye. This temperature distribution information is helpful for identifying and measuring occluded targets in the field of security, and determining the location and status of targets through temperature differences.

Many scholars at home and abroad have proposed various reconstruction methods for infrared 3D reconstruction. In 2020, Sabato et al. verified the feasibility of using SFM technology for 3D reconstruction based on infrared images through experiments[3]. However, the author could only establish a 3D grayscale model and could not directly or indirectly read temperature information in the 3D grayscale model. Subsequently, Zheng Haichao[4] in 2022 and CAI Hongbin et al. [5] in 2023 respectively conducted research on the thermal infrared 3D model reconstruction method of infrared images obtained by unmanned aerial vehicle aerial photography based on SFM technology, and attempted to recover temperature information from the reconstructed 3D gray model. However, for a long time in the past, due to the limitations of thermal imaging detection and sensing technology, the resolution of infrared detectors was generally low and lacked high-frequency features. The number of point clouds reconstructed by traditional methods was scarce, and the numerical and positional errors of the generated points were large. Especially in situations where the environment lacked identification and objects were symmetrical, there was even a phenomenon where 3D reconstruction could not be performed.

One solution to the above problem is to use information from different modalities captured at the same location to assist in infrared 3D reconstruction. Due to the different information that different modalities focus on and the varying imaging accuracy between different modality cameras, it is common in engineering to use one camera to capture target images for localization or recognition tasks, and then assist images from other modalities to further complete more refined visual tasks[6-8]. The common approach for 3D reconstruction is to first capture images with a regular camera and complete the 3D reconstruction, then match the image information of the infrared camera to the point cloud[9-11], or use other methods such as structured light[12-13], TOF depth data[14-15], laser point cloud[16], etc. to register the combined thermal infrared images. The advantage of this scheme is that it can obtain a dense three-dimensional point cloud of the target object, and then map the temperature information of the point cloud to obtain a more accurate three-dimensional temperature field distribution. In recent years, the research focus of this scheme has been on improving point cloud matching and positioning accuracy[17-19], as well as reducing the impact of thermal infrared imaging defects [20]. Although the above method effectively solves the sparsity of 3D reconstructed point clouds and improves positioning accuracy, it requires high equipment requirements and generally requires high-definition thermal infrared cameras to be used in conjunction with other specially designed cameras. And the aforementioned 3D reconstruction techniques rely on explicit expression, and the improvement of accuracy will occupy a large amount of storage space, so a trade-off must be made between the two.

In addition to joint multimodal assisted passive infrared reconstruction, Marie-Marthe Groz et al. also proposed using active infrared imaging to measure the temperature field of surface measurements to characterize three-dimensional heat sources buried in materials[21]. However, this method has poor generalization and does not have universality, making it difficult to promote and apply on a large scale. Another emerging implicit 3D reconstruction method based on neural networks in recent years is the neural radiation field 3D reconstruction technique, which greatly improves the quality of image reconstruction compared to traditional methods. This method expresses the discrete three-dimensional voxel information in space as a continuous function. The model uses an MLP network to fit this function and samples the voxel information on the corresponding light rays of a given pixel in this function, and renders the given pixel information through volume rendering methods. This method does not rely on SFM reconstructed point clouds for 3D reconstruction, but directly learns scene information from the image, avoiding the impact of SFM on sparse and large errors in infrared scene reconstruction point clouds. In addition, the three-dimensional scene information reconstructed by neural radiation field is implicitly expressed and stored in the weight coefficients of the

neural network, which ensures that the improvement of scene reconstruction quality does not consume more storage space and is more convenient for transmission and storage. It has important applications in remote sensing and other fields. However, as an algorithm that relies on neural networks, the neural radiation field has problems with slow inference speed and long convergence time. Some effective improvements, such as PointNeRF[23], Plenoxels[24], and 3D-GS[25], reduce the role of MLP in voxel inference by introducing more point cloud data. This not only improves speed but also gradually degrades the algorithm back to traditional methods that rely on point clouds and consume high storage, making it difficult to meet practical application scenarios with limited storage. In practice, the results reconstructed by SFM for general scenes captured by ordinary thermal infrared cameras are difficult to meet the requirements of the above algorithms for initial point clouds, and randomly generated initial point clouds are difficult to support accurate reconstruction.

In response to the current situation of low resolution thermal infrared 3D reconstruction, firstly, thermal infrared cameras generally have fewer pixels, resulting in low infrared image resolution and fewer high-frequency features, making it difficult to reconstruct through SFM 3D. This paper chooses to use neural radiation field technology to learn scene information from images, breaking away from the dependence of traditional methods on feature point clouds; Secondly, traditional methods have a huge consumption of storage space. In this paper, we use MLP network to implicitly express the 3D voxel information of the scene, ensuring that the reconstruction accuracy is improved while reducing the storage space occupation. In addition, in order to solve the problem of slow rendering speed and accelerate the inference and convergence speed of neural networks. Based on the fact that infrared information mainly focuses on low frequencies, this article prunes the network layers and reduces the network size without compromising image quality. And a semantic segmentation network was introduced to screen and process the training data, achieving network acceleration. Finally, this article improved the network structure of the neural radiation field based on the principle of thermal infrared imaging, making it more suitable for the low resolution and temperature expression characteristics of infrared images, further improving rendering accuracy, and achieving high-quality reconstruction at extremely low resolutions.

1 Infrared NeRF network

Inspired by the concept of light field in physics, the core of NeRF proposed by Mildenhall et al. (hereinafter referred to as the original NERF) is to simplify the light field model. Based on the characteristic that the color and radiation intensity of visible light do not change significantly with distance, a five dimensional directional vector and an RGB three-dimensional color vector are used in the original NeRF to represent all the information of a light ray. However, in infrared imaging scenarios, this simplified approach no longer conforms to thermodynamic models. Because the false color presented in common thermal infrared images is not characterized by wavelength differences, but by the intensity of infrared radiation energy received by the infrared detector to determine the depth of color, it is called "false color". Generally speaking, thermal infrared images are grayscale images, and areas with higher grayscale values receive higher radiation intensity. Unlike visible light, infrared radiation intensity has a quadratic inverse relationship with distance, which means that even if the same object is photographed from the same perspective, the grayscale (false color) of the image will significantly change with the change of shooting distance. In the original NERF, if the viewing angle is fixed, the color of the light is fixed, that is, it is not affected by distance. In thermal imaging, changes in distance can lead to variations in grayscale (false color). If the original NeRF network is used, it means that inputs from the same perspective but different distances may correspond to significantly different outputs, which is unfavorable for 3D modeling. So this article considers incorporating the changes in distance into the original NERF model. Not only considering the position and observation direction of the sampled voxel, but also taking into account the distance (observation distance) between the imaging point of this sampling and the sampled voxel.

Different from the 5D input of the original NERF, this paper represents continuous scenes as 6D vector valued functions, whose inputs include 3D position, 2D observation direction, and 1D observation distance, and whose outputs are grayscale feature values and volume density. Similar to the original NeRF, this article uses a three-dimensional Cartesian coordinate vector instead of the observation direction. Then use an MLP network to approximate the continuous 6D scene, also known as implicit representation, iteratively optimize its weights, and map each input 6D coordinate to its corresponding volume density and grayscale feature value.

Considering that volumetric density is also an intrinsic property of the target object in infrared scenes and does not change with the observation distance and direction, this paper adopts the same method as the original NeRF to restrict the network from predicting volumetric density σ as a function only related to position X; And predict the grayscale value g as a function related to position, observation direction, and observation distance, in order to encourage the model to be sensitive to both observation direction and distance.

Compared to visible light imaging with RGB three channels, infrared imaging only contains one-dimensional grayscale values. As shown in Figure 1 (b) and Figure 1 (c), due to the fact that thermal infrared images usually contain more low-frequency information, the information in the spectrum is mainly concentrated in the low-frequency part, and the information entropy is relatively small. Therefore, using the parameter scale of the original NERF model that expresses visible light scenes to characterize the redundancy phenomenon in hot red

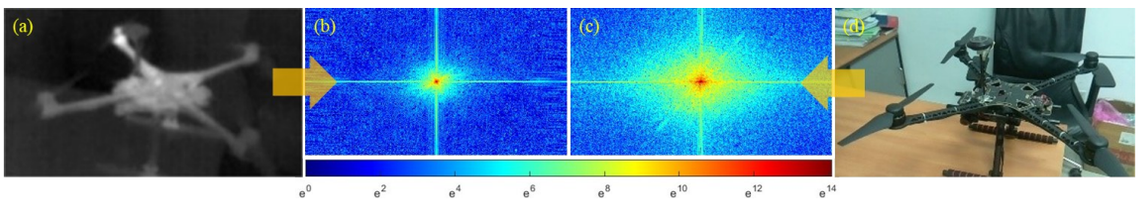

Fig 1 Imaging effect and spectral distribution of different wavebands: (a) Infrared image; (b) Spectral distribution of infrared image; (c) Visible distribution of infrared image (d) Visible image;
图1 不同波段的成像效果和频谱分布图

scenes of the same scene. Not only will it slow down the inference speed of the model, but it will also increase the difficulty of model convergence. Therefore, this article chooses to compress the number of layers in the original model and reduce the dimensionality of the position encoding function to significantly improve the inference speed of the model without compromising imaging quality.

The MLP network first processes the input 3D coordinate X using 6 fully connected layers (activated with ReLU, each layer having 256 channels), and outputs density σ and 256 dimensional feature vectors. Then connect the feature vector with the observation direction of the camera light and pass it to another fully connected layer (using ReLU activation and 256 channels), connecting the output feature vector of this layer with the observation distance. Finally, it is passed to another fully connected layer (activated with ReLU, 256 channels), which outputs grayscale values related to the viewing angle and distance. At the same time, this article optimizes the two methods of position encoding and hierarchical volume sampling in the original NeRF. The position encoding function is shown in formula (1):

$$\gamma(p) = \left(\sin\left(2^0\pi p\right),\cos\left(2^0\pi p\right),\cdots,\sin\left(2^{L-1}\pi p\right),\cos\left(2^{L-1}\pi p\right)\right) \tag{1}$$

Formula (1) is applied to the three coordinate values in X (normalized to [-1, 1]), the Cartesian observation direction unit vector $d(\theta,\varphi)$ (normalized to [-1, 1]), and the observation distance l (normalized to [0, 1]). In this experiment, set L=8 for γ(X), L=4 for γ(d), and L=8 for $\gamma(l)$. Related experiments on the selection of L-value combinations are presented in Model Efficiency Optimization. For layered volume sampling, this article also uses two networks (coarse network and fine network) to improve rendering efficiency. The output of the coarse network is used to estimate which points on the entire ray are more likely to be on the surface of the object. Based on the estimation results, the ray is resampled and further fed into the fine network for inference

2 Infrared data screening methods

In real production and engineering infrared 3D scene modeling applications, more attention is usually paid to high/low temperature objects in the foreground rather than the changes in the background information that are relatively homogeneous in temperature. Therefore, if it is desired to perform fine 3D reconstruction of the target object in the foreground instead of reproducing the details of the background (which usually cannot be rendered in thermal IR images), it is not necessary to train and render the rays of all input pixels.

As shown in Fig. 2, there is almost no pixel variation in the background region in infrared pictures, so the information of a region containing several pixels can be approximated by the information of one of the pixels or the average value of the region. The pixels in the foreground region, on the other hand, change more drastically, and thus attention must be paid to the information provided by each pixel. For pixels in the foreground region, the algorithm must learn the 3D scene based entirely on the original information and add all the ray information to the training data on a pixel-by-pixel basis; for the background region the algorithm simply replaces the information of some pixels with the average information of a number of pixels or reduces the frequency of rays occurrences in a piece of similar area, with the goal of trying to reduce the number of invalid rays and thus reduce the amount of training data fed to the network.

As shown in Fig. 3, in order to filter out the effective light that is really used for training, it is first necessary to obtain each image in the target scene and the corresponding camera position, in this paper, the original image is inputted into a Semantic Segmentation Network (SSN) to get the mask image of each image, which is a binary-valued image, and the foreground and background in the image can be segmented by this mask. The algorithm then calculates the photocenter coordinates (Pos) and the direction of rays(Dir) in the world coordinate system for each pixel in the image based on the camera pose of each image, and concatenate them together with the gray-scale value (Gray) to form vectors. The ray vectors generated for each image are divided into foreground and background according to their corresponding positions on the mask image, and then all the foreground ray vectors are stacked into a matrix, while the remaining background ray vectors are constrained to appear at a frequency of 1/n in the training data. During model training, for each ray vector of the coarse network, Nc sampling points are randomly selected on the ray according to its photocenter coordinates (Pos) and ray direction (Dir), and the position encoding is performed by applying formula (1) to the coordinates (Position), direction (Dir), and depth of each sampling point, and the encoding matrix is finally spliced with the gray-scale value

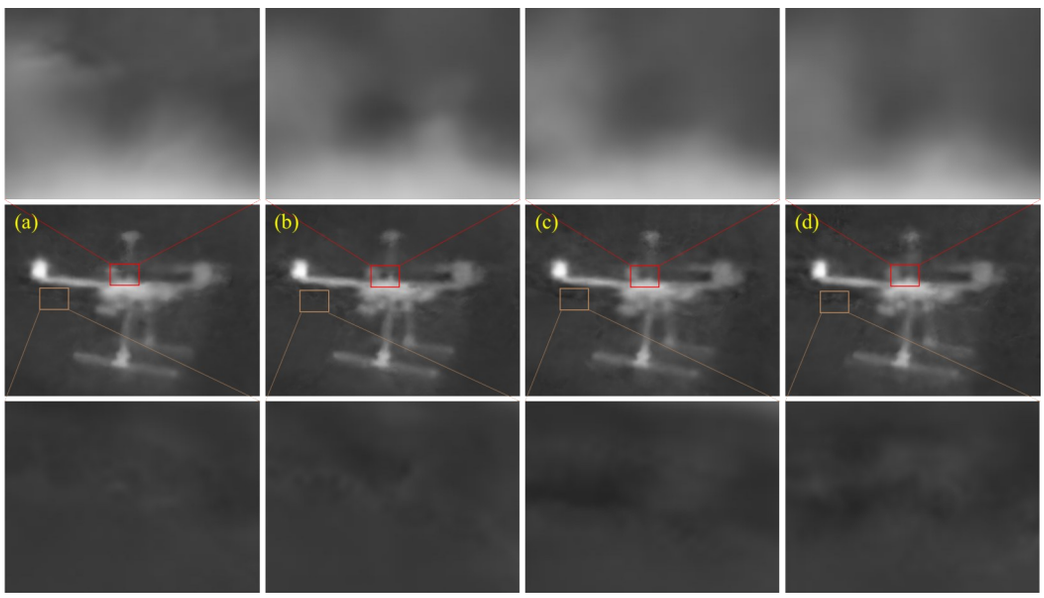

Fig 2 Reconstruction effect of different background pixel groupings: (a) Two groups; (b) Four groups; (c) Six groups; (d) Nine groups
图2 不同背景像素出现频率的重建效果

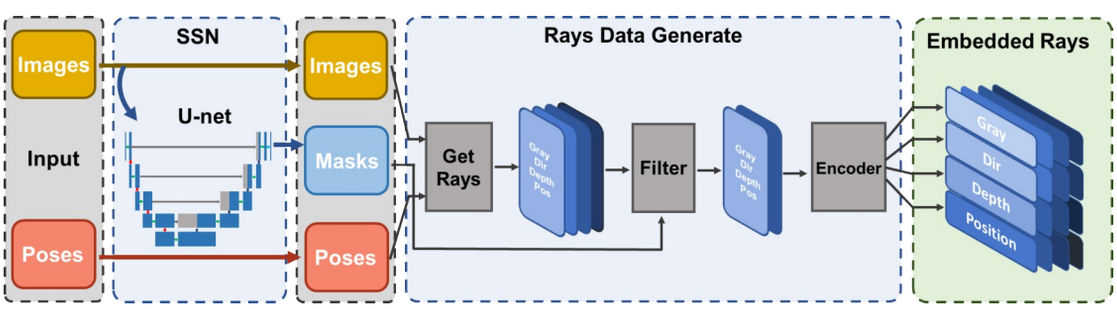

Fig 3 Split and process to get coded training data
图3 分割并处理得到编码训练数据

(Gray) to obtain the data matrix of a ray, and stacked with other ray data in the same batch to form the encoded training data (Embedded Rays) of that batch. Finally, the encoded matrix is spliced with the gray-scale value (Gray) to obtain the data matrix of a ray, and stacked with other ray data in the same batch to form the encoded training data (Embedded Rays) of the batch. The process of generating training data for the fine network differs only in the way the sampling points are selected, and the process is similar to the coarse network.

Given that convolutional networks also have excellent performance in various classification tasks, this paper chooses to semantically segment the input image based on convolutional networks. Under the premise of certain network depth and small samples, the model is made to maintain good segmentation performance without overfitting. Considering the accuracy and speed of the network, this paper chooses the U-Net network, which has excellent performance in medical gray-scale image segmentation task, as the target network. In the experiment, 15 images in the dataset are manually labeled and the training of the U-Net network is completed, and then all the images in the NeRF training set are applied to the above trained U-Net model for semantic segmentation. The experimental segmentation results are shown in Fig. 4, Fig. 4(a~c) shows the unsegmented infrared imaging, while Fig. 4(d~f) shows the results after U-Net segmentation. It can be seen that U-Net has a good segmentation effect on the dataset and accurately captures the boundaries of the foreground and background images on both untrained datasets (Fig. 4(b) and Fig. 4(c)), but nulling occurs in the regions where the pixels inside the foreground objects change drastically. For the void problem, this paper adopts the contour filling technique to find out the void contour in the image and then the void region is filled by flooding method. The filling results are shown in Fig. 4(g~i), which shows that the filled mask image not only covers the foreground target, but also does not occupy extra background. Experiments show that the segmentation ability of the U-Net model trained with small samples for front and back views in infrared scenes has the potential to be generalized to general scenes. For specific scenes, a small number of images can be labeled for U-Net training to achieve good segmentation results. For a wider range of scenes, data segmentation can also be achieved by applying the U-Net segmentation results to fill in the voids and then synthesize the masks.

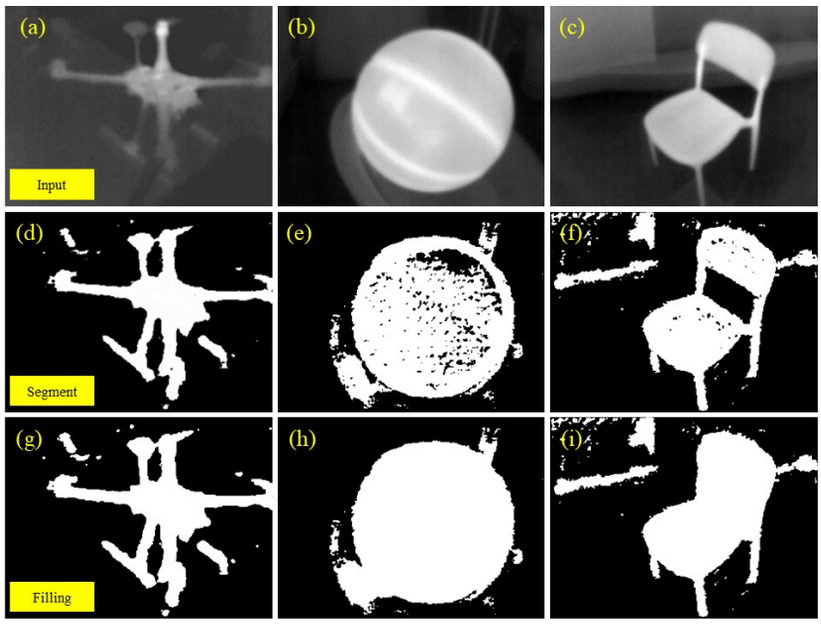

Fig 4 Segmentation test of infrared imaging scene by U-Net
图4 U-Net对红外成像场景的分割试验

3 implementation and Comparisons

3.1 Dual-Band Assisted Camera Pose Transferring

The data fed to the model for training in the experimental process include the pixel information of the image and the camera pose corresponding to the image. For the camera pose, this paper uses colmap to solve the camera pose after obtaining the infrared image. During the experiment, it is found that colmap is not able to obtain the thermal infrared camera pose with low resolution infrared images, For example, by directly applying colmap to one of the 210 images in the infrared dataset, only 13 images can be successfully matched to generate 106 points. Ideally, a dataset containing 461 images would only have a matching success rate of about 65% of the matching success rate, and the 298 images that were successfully matched would only generate a total of 1208 points. It can be seen that it is difficult to accurately calculate the camera position just from the captured thermal infrared images, and some of these experimental results are shown in Fig. 5(d) and Fig. 5(e). In the experiment, the images of two wavebands were taken by binocular camera around the UAV, in which Fig. 5(e) shows the calculation results of colmap in infrared band, which can be seen to have more matching error points, and Fig. 5(d) in order to obtain the reference reconstruction results for visible bands (considered as the true value).

Aiming at the above problems, this paper proposes a dual-band assisted camera position transfer calculation method, and the main idea is shown in Fig. 5 above. It can be seen that although the thermal infrared scene cannot be directly reconstructed by SFM, it can well calculate the camera pose of the RGB scene, and since the pose relationship between the infrared camera and the

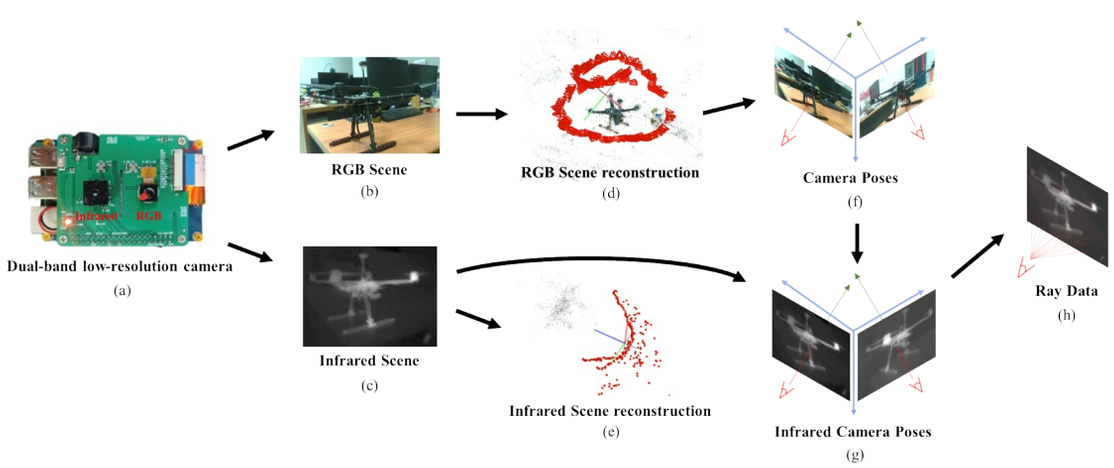

Fig 5 Bimodal assisted calculation of light information
图5 双波段辅助相机位姿转移计算原理图

visible camera can be accurately calibrated[8], the visible camera's pose can be transferred to the infrared camera through the rotated matrix, so that the pose of the infrared camera can be indirectly estimated by the visible camera's pose.

As shown in Fig. 5, this paper designs a process of acquiring ray position information for bimodal assisted computation. Figure 5(a) shows an integrated camera with infrared and visible light detectors, and because the baselines of the two detectors are small, it can be approximated that the data acquired from the two modalities are taken at the same location. Therefore, the SFM technique is applied to the visible image to obtain the camera pose of each image in the world coordinate system, and then the camera pose is matched with the infrared image taken at the same location of the corresponding visible image, based on which the ray pose information is calculated.

3.2 Image Segmentation and Minimum Sphere

Inference speedup in NeRF has been a hot topic of academic discussion, and the most direct way to improve inference speedup is to reduce the number of rays rendered. To address the problem of inference speed enhancement this paper unfolds a non-uniform spatial ray optimization method using image segmentation, the ray information generated by the foreground pixels in the training set after U-Net segmentation is fully incorporated into the training data, and in order to study how much of the background rays are retained, as well as how to study how to retain the pixel information in the background can be maximized to minimize the amount of data without affecting the quality of the scene reconstruction. In this paper, experiments were carried out in which the pixels of the background were divided into two groups (every two neighboring pixels were included in a different group), four groups (every 2*2 pixel grid pixel was included in a different group), six groups (every 2*3 pixel grid pixel was included in a different group) and nine groups (every 3*3 pixel grid pixel was included in a different group). Therefore, the final ray for each batch of training data is the combination of one of the background rays and all foreground rays, traversing all the above combinations by batch. In other words, the pixel information of the background rays occurs at the frequency of 1/2, 1/4, 1/6 and 1/9, respectively.

As shown in Fig. 2, the experimental results of this paper are mainly reflected in two aspects, which are the reconstruction quality of the foreground target data (red area in Fig. 2) and the scene background (orange area in Fig. 2). As the number of groups increases, the reconstruction quality of the internal details of the target object is improved, and the local details of the target gradually become clear and highlighted. However, more meaningless noise is visible in the background, which means the reconstruction quality of the background decreases. And the demarcation between the target object and the background gradually becomes blurred. The experimental results show that while the overall image quality decreases when dividing pixels into four groups, the training time is only about 25% of the time for full retention, and continuing to segment pixels does not result in a significant speedup but reduces the rendering quality.

In fact, experiments have revealed that many real-world infrared 3D scene reconstruction tasks do not include the reconstruction of the scene, i. e., the background of the image, and that many thermal infrared images essentially contain no effective background information. At this point, if we consider only the 3D reconstruc-

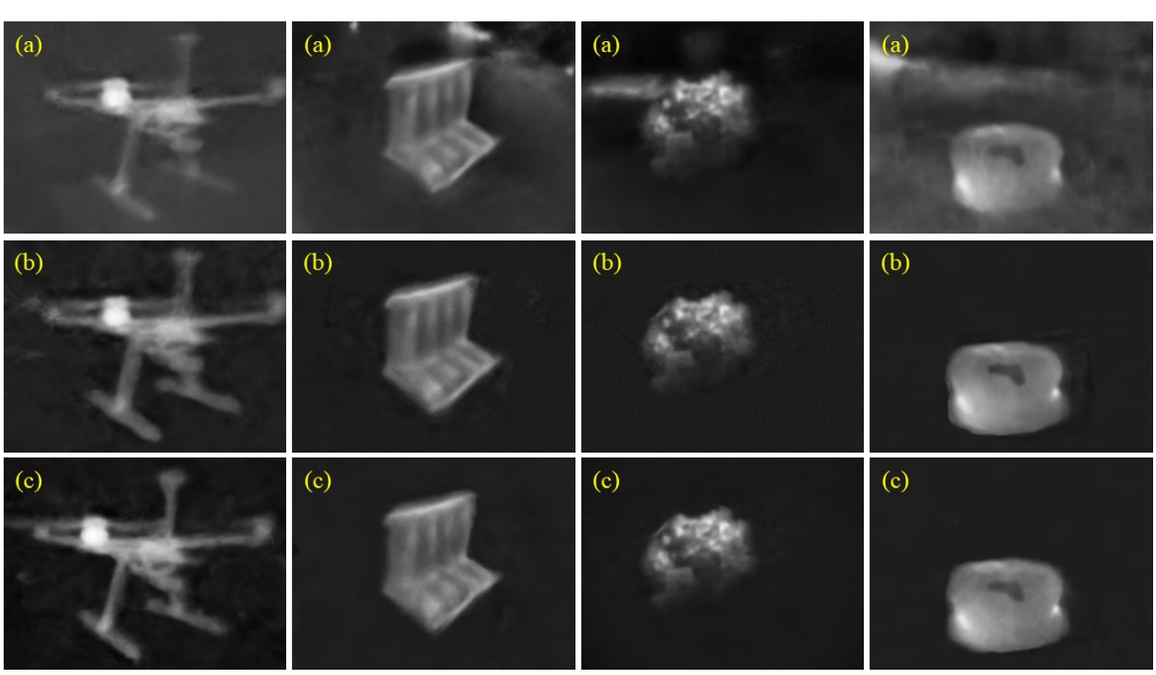

Fig 6 Reconstruction effects of different acceleration strategies: (a) Background colour is preserved; (b) The background is covered in black; (b) Narrowing the Query Interval with Minimum Coverage Balls
图6 不同加速策略所对应的重建效果

tion of the target object in the foreground and discard the background information, the model can theoretically be allowed to focus entirely on learning the target object, achieving faster training and better reconstruction quality. In this paper, we first estimate the average gray value of the background in the thermal infrared scene to be about 30. To improve the discrimination between target and background and prevent neural networks from being unable to train due to extreme values (0), we set the pixel gray value of all backgrounds to 30 and select the central pixel from each 3x3 pixel grid as the target ray to be included in the training data. The experimental results, shown in Figure 6, show that the inference time has been improved without compromising the quality of the reconstruction of the target object, consuming only about 20% of the time required for full-image input. In addition, a certain grey level difference between the background and the target objects makes it easier to visually distinguish the details and edges of the target objects.

Based on the camera poses input into the algorithm, each camera's frustum can be represented by four rays, as shown in Figure 7(a). Assuming that the target object is always within the image during shooting, and that the object is always within the frustum of each camera in 3D space, the object must be within the intersection of all frustums. Through experimental observation, it can be intuitively seen that a frustum passing through (intersecting with) the sphere is equivalent to all four rays of the frustum intersecting with the sphere. In this paper, a minimum sphere that is tangent to or intersects all frustum rays is computed using the simulated annealing method. As shown in Figure 7(b), such a sphere intersects with all frustums and thus encompasses the intersection area of all frustums, covering the target 3D object. The final reconstruction result is shown in Figure 6(c), where it is evident that the image quality has been significantly improved, especially at the boundaries between the target and the background.

This design ensures that most coarse network queries are not computed by the network but directly return a zero value, while the fine network queries, after inverse transformation, are more concentrated within the aforementioned sphere. This is equivalent to compressing the sampling interval and increasing the sampling frequency. Therefore, this paper hypothesizes that it is possible to appropriately reduce the number of samples on the fine network, allowing sampling at a frequency similar to the original method within the compressed sampling interval, thereby further improving inference speed while maintaining image quality. As shown in Table 1, just using the minimum sphere to reduce the query interval can increase the inference speed to 2.36 times. When the fine sampling frequency is reduced on the basis of the minimum sphere, not only is the inference speed greatly improved, but the reconstruction metrics also show improvement compared to not reducing the sampling frequency. When the fine sampling frequency is reduced to 0.5 times, the metrics improve significantly and the inference speed increases to 3.43 times that of the original method; when the fine sampling frequency is reduced to 0.25 times, the metrics are comparable to those of the minimum sphere only and the inference speed increases to 4.69 times that of the original method.

3.3 Algorithm Evaluation and Comparison

The original Neural Radiance Field (NeRF) did not consider the influence of the distance between the sampling point and the imaging point on the imaging. To compare the advantages of the proposed algorithm with

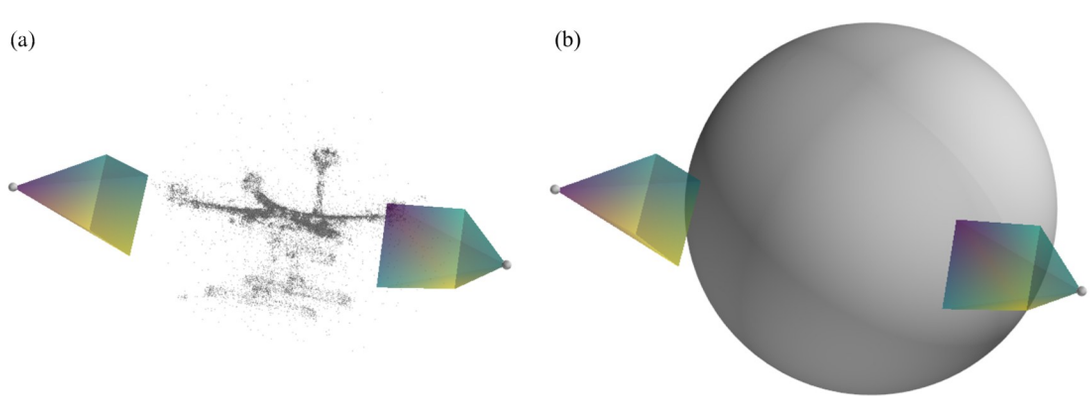

Fig 7 Schematic diagram of the algorithm to compute the intersection of the minimum sphere covering the optic cones: (a) Two view cones and the target object in them; (b) Minimum coverage sphere covering the intersection of the optic cones
图7 计算最小球覆盖视锥体交集的算法示意

Table 1 Reconstruction speed and metrics for different model structures
表1 不同模型结构的重建速度与指标

	Speed(kRPS)	PSNR	SSIM	LPIPS
the original method	4.98	32.45	0.924	0.081
the minimum enclosing sphere method	11.77	29.74	0.885	0.243
the minimum enclosing sphere downsampling method 1/2	17.10	30.59	0.904	0.219
the minimum enclosing sphere downsampling method 1/4	23.34	29.83	0.895	0.239

previous algorithms, and to verify the role of query direction and distance information in the NeRF reconstruction process, this paper uses the controlled variable method to design experiments with combinations of adding or not adding direction and distance information, while keeping the model and other structures unchanged. The training data and data screening methods are the same in all experiments. The experimental results are presented in Tables 2 to 4.

Table 2 PSNR of different model structures for different scene reconstructions
表2 不同模型结构对不同场景重建的PSNR

	Plane	Cup	Flower	Pillow	Fan	Bottle	Tower
MLP	18.10	25.08	27.90	30.31	28.67	24.27	23.45
Depth	19.02	28.07	30.63	32.05	30.15	25.60	25.45
NeRF	25.62	32.54	38.68	36.59	37.38	28.78	33.47
3D-GS	25.68	32.93	38.36	37.03	36.83	28.30	33.00
Infrared NeRF	**30.05**	**37.84**	**41.58**	**40.48**	**40.38**	**35.82**	**37.65**

Table 3 SSIM of different model structures for different scene reconstructions
表3 不同模型结构对不同场景重建的SSIM

	Plane	Cup	Flower	Pillow	Fan	Bottle	Tower
MLP	0.646	0.871	0.911	0.936	0.928	0.878	0.865
Depth	0.683	0.904	0.933	0.944	0.942	0.893	0.897
NeRF	0.795	0.942	0.973	0.966	0.970	0.915	0.949
3D-GS	0.878	0.966	0.985	0.978	0.986	0.953	0.976
Infrared NeRF	**0.908**	**0.976**	**0.989**	**0.981**	**0.989**	**0.963**	**0.988**

Table 4 LPIPS of different model structures for different scene reconstruction
表4 不同模型结构对不同场景重建的LPIPS

	Plane	Cup	Flower	Pillow	Fan	Bottle	Tower
MLP	0.360	0.203	0.142	0.141	0.138	0.228	0.241
Depth	0.349	0.165	0.114	0.126	0.110	0.203	0.190
NeRF	0.342	0.119	0.054	0.091	0.066	0.162	0.097
3D-GS	0.222	0.076	0.024	0.061	0.028	0.108	0.034
Infrared NeRF	**0.220**	**0.062**	**0.020**	**0.060**	**0.022**	**0.096**	**0.014**

As shown in Tables 2, 3 and 4, the results indicate that whether direction or distance information is added, the reconstruction capabilities of the model, as evaluated by metrics such as PSNR, SSIM and LPIPS, are significantly improved compared to when neither is added. It is evident that the algorithm in this paper has a notable enhancement effect and outperforms existing NeRF models that only consider directional information in terms of scene reconstruction capabilities. In addition, it has a considerable advantage of a smaller model size compared to the 3D-GS method, which explicitly represents voxel fields.

As shown in Figure 8, in addition to quantitative evaluations, this paper also conducted qualitative assessments of the reconstruction quality of different models on various datasets. It can be observed that images reconstructed solely using MLP exhibit numerous artifacts and blurred details. While the original NeRF provides clearer reconstruction details, artifacts are still present. The inclusion of query distance alone reduces image artifacts but does not provide sufficiently clear reconstruction details. The 3D-GS method, which leverages point clouds for reconstruction, produces images with clear details but suffers from partial void phenomena, such as the antenna in Figure 8 (Plane) and the center of Figure 8 (Bottle). However, the method proposed in this paper improves both reconstruction details and image artifacts.

In addition, as shown in Figure 9, our algorithm significantly outperforms the original NeRF algorithm in terms of convergence speed. During the initial training phase, the loss value of our algorithm decreases much faster than that of the original NeRF. This indicates that our algorithm can find better parameters more quickly in the early stages of training, leading to a rapid reduction in the loss value. By the time the training reaches 20,000 iterations, the loss value of our algorithm is already lower than the loss value of the original NeRF algorithm at the end of its training (100,000 iterations), and it continues to decline slowly, maintaining a very low loss value level, with a decline rate still higher than that of the original NeRF.

3.4 Model Efficiency Optimization

As previously mentioned, the frequency information in thermal infrared scenes is often lower than that in visible light images within the same scene. Considering that MLP networks have good extraction capabilities for low-frequency features, and to further accelerate model inference speed, this paper aims to reduce the size of the MLP without compromising reconstruction quality. This paper gradually reduced the number of network layers from the original eight-layer algorithm to two layers and conducted experiments at each step. The network structures used in all experiments adhered to the design proposed in this paper, and the same training data and data filtering methods were employed. The experimental results are shown in Figure 9. The PSNR values gradually increase as the number of layers ranges from two to five, peaking at six layers, and then decrease with further increases or decreases in layer count. SSIM exhibits a similar trend to PSNR. In contrast, LPIPS gradually decreases as the number of model layers increases, with a significant drop between five and six layers, followed by more gradual decreases at other times. In summary, for thermal infrared image models, the sixth layer offers the best overall performance.

In this paper, experiments are conducted for the selection of the L-value combination in formula (1), and the experimental results are shown in Table 5, where the 848 combination performs the best in both PSNR and SSIM key metrics, and the LPIPS value is at a low level. Therefore, this combination achieves a good balance between signal fidelity, visual quality, and perceptual consistency, and is suitable as the final choice.

In all experiments conducted in this paper, the other training parameters were used as below: 1024 rays

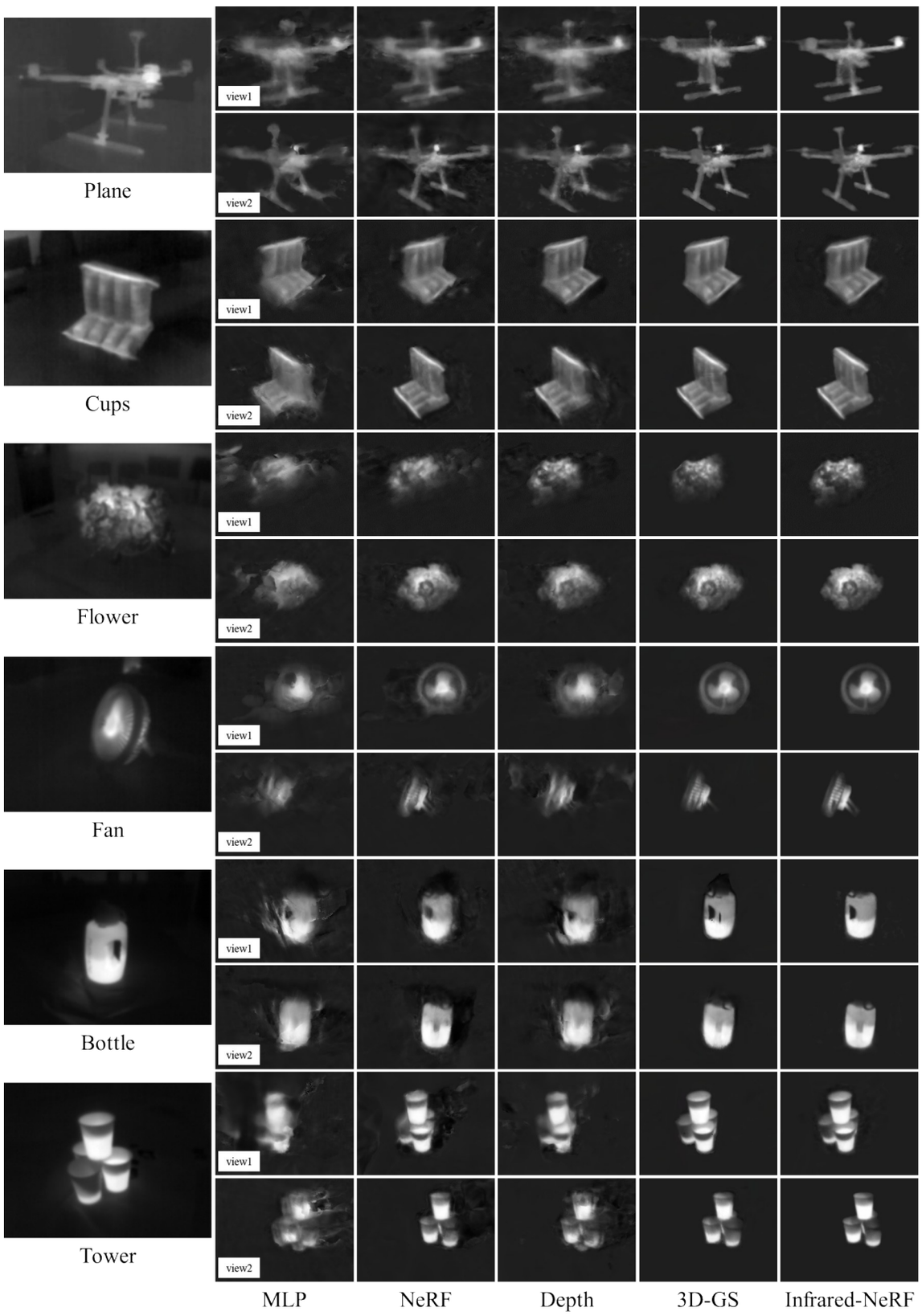

Fig 8 Reconstruction effects of different model structures:(a) MLP only; (b) Add enquiry direction only(NeRF); (c) Add enquiry distance only; (d) 3D-GS; (e) Add enquiry direction and enquiry distance(Ours)
图8 不同模型结构的重建效果

were processed per training iteration. After generating each ray, 64 coordinates were sampled along the ray and

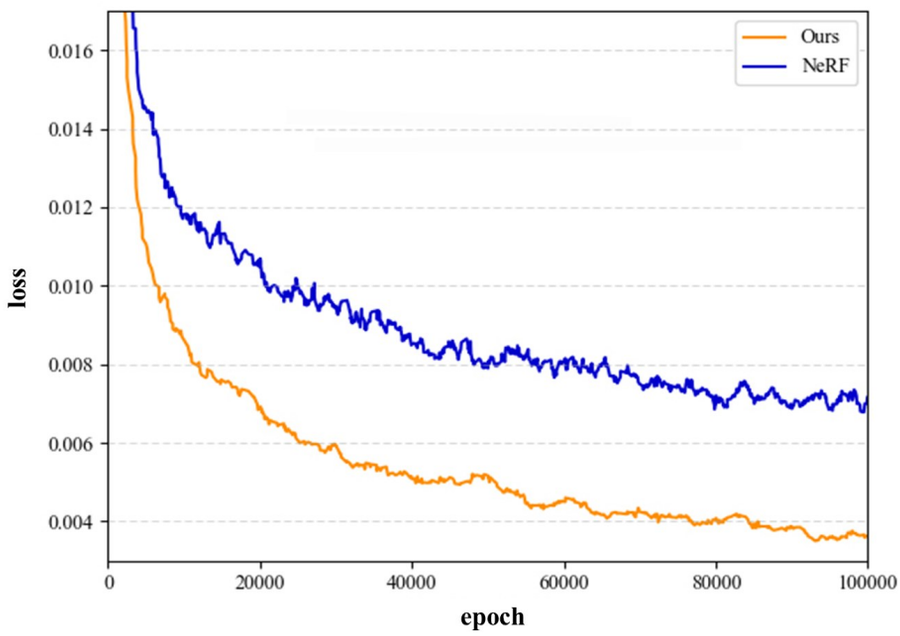

Fig 9 Comparison of Loss Values at Different Training Iterations

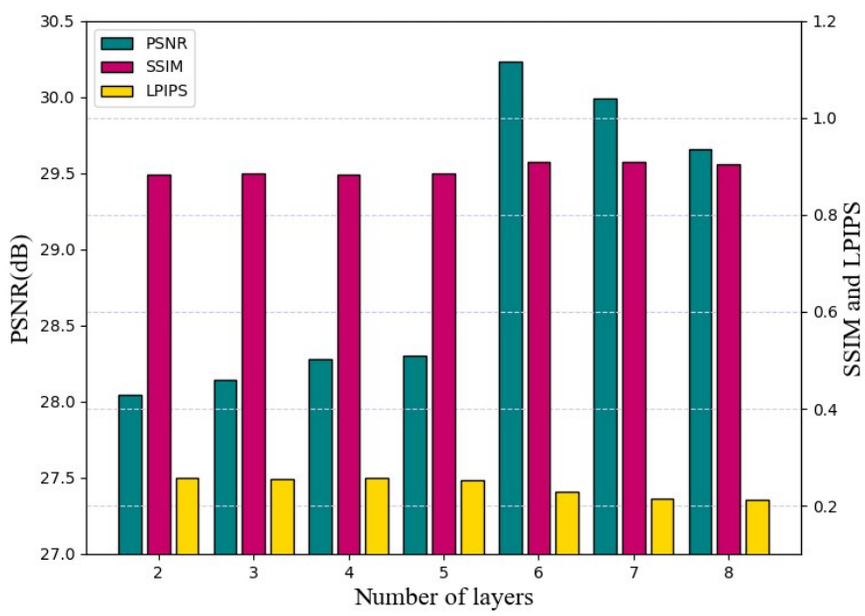

Fig 10 Reconstruction results for different model layers: (a) The left axis identifies the value of PSNR and the right axis identifies the values of SSIM and LPIPS
图 10 不同模型层数的重建效果

fed into the coarse network for sampling (Nc=64), and another 64 coordinates were sampled along the ray and fed into the fine network for sampling (Nf=64). The Adam optimizer was used in the experiments, with the learning rate starting at 5×10^{-4} and decaying exponentially to 5×10^{-5}. The other Adam hyperparameters were set to their default values: $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\varepsilon = 10^{-7}$.

4 Conclusion

This study addresses the challenges of 3D reconstruction in thermal infrared scenes by leveraging neural radiance field (NeRF) technology and exploring five technical approaches to achieve realistic reconstruction results. By exploiting the inherent relationships between thermal infrared scene image data and query distances, the network framework was redesigned with optimized parameter scales. Notably, the hidden layers were reduced to six, enhancing both reconstruction quality and inference speed. Experimental results demonstrate that the improved network achieves an average increase of 13.8% in PSNR and 4.62% in SSIM, while reducing LPIPS by 46%.

Table 5 Reconstruction results of different L-value combinations
表5 不同L值组合的重建效果

$L_x L_d L_l$	PSNR	SSIM	LPIPS
228	27.2953	0.8721	0.2952
328	27.6051	0.8786	0.2848
428	27.8444	0.8810	0.2699
528	28.0037	0.8835	0.2612
628	28.0763	0.8853	0.2566
728	28.0118	0.8828	0.2593
818	27.2522	0.8730	0.2590
822	27.9878	0.8833	0.2572
823	27.9288	0.8831	0.2570
824	27.8735	0.8824	0.2573
825	28.0167	0.8837	0.2575
826	27.8967	0.8824	0.2605
827	28.0729	0.8840	0.2617
828	28.1121	0.8850	0.2553
838	28.9761	0.8946	0.2621
848	**29.9096**	**0.9053**	**0.2647**
858	29.4527	0.9006	0.2462
868	29.4320	0.9004	0.2472

To address the slow inference speed inherent in the original NeRF, the study takes advantage of the characteristics of thermal infrared scenes by incorporating a semantic segmentation network to preprocess the training data. Foreground and background segmentation is performed, and contour-filling techniques are used to handle mask data holes. By reducing the weight of background data, the NeRF learns to reconstruct thermal infrared scenes more efficiently, converging in only 25% of the time required by the original NeRF.

Additionally, to further enhance inference speed, the study proposes an algorithm for calculating the minimum bounding sphere that covers the intersection of the visual cone. This reduces the query space significantly, increasing inference speed by 4-6 times compared to the original NeRF while maintaining reconstruction quality. These advancements provide an efficient and high-quality solution for 3D reconstruction in thermal infrared scenes.

5 Discussion

In all thermal infrared scenes, our method outperforms the current state-of-the-art (SOTA) approaches for 3D reconstruction (NeRF and 3D-GS). By introducing a hidden layer that accounts for distance into the NeRF model, we enable the model to learn the physical property of thermal infrared radiation attenuation with increasing distance, thereby improving the reconstruction quality of thermal infrared scenes.

Point cloud-based methods, such as 3D-GS, face limitations in reconstructing thermal infrared scenes because of the low information density inherent in thermal infrared images. Traditional point cloud reconstruction techniques struggle to generate initial point clouds from thermal infrared data. 3D-GS starts from random initial point clouds, often failing to accurately reconstruct object boundaries and resulting in gaps in the reconstruction.

The demand for 3D reconstruction of thermal infrared modalities is substantial in practical applications. For example, in the construction industry, there is a need to obtain 3D thermal structures of buildings for heat effect analysis; in medicine, 3D thermal structures of the human body are needed to analyze pathological changes; in remote sensing and autonomous driving, 3D thermal structures of targets are required for further downstream tasks, and so on. We found that these tasks share a common characteristic: they do not require the reconstruction of the entire thermal infrared scene, but rather the reconstruction of a key object within the scene. In fact, while the main object occupies the majority of the scene's information, it often constitutes only a small portion of the image. Thus, we propose a method that segments the scene and assigns different weights to different rays, allowing the model to learn the main object more quickly and accelerating the convergence of model training. Furthermore, by using a minimum bounding sphere to limit the sampling boundary, we significantly improve the model's inference speed.

Compared to other methods, our approach makes significant progress in offline reconstruction tasks for static thermal infrared scenes, making it suitable for applications in fields such as construction and medicine. However, our method still involves trade-offs when compared to other approaches. The reconstruction of all scenes with our method takes at least 2-3 hours, whereas 3D-GS can reconstruct a scene in less than 30 minutes. However, due to its point cloud-based nature, 3D-GS often results in blurry edges and gaps in the reconstructed scene and requires more storage space for point cloud data. In contrast, our method generates scenes with higher reconstruction quality according to all metrics, as well as human visual assessment, while requiring less storage space for network parameters.

Acknowledgment

This research was support by The Fundamental Research Funds for the Central Universities (Integrated Innovation Category 2024300443) and the National Natural Science Foundation of China (NSFC) Young Scientists Fund (Grants No 62405131).

References

[1] Y. Sun, Research on the Application of City Information (CIM) Platform in Smart Cities, Wisdom China, 2022.
[2] B. Chen, "Application of Infrared Thermal Imager in Medicine and Research on Measurement", Qinhuangdao, Yanshan University, 2010.
[3] A. Sabato, M. Puliti, C. Niezrecki, "Combined infrared imaging and structure from motion approach for building thermal energy efficiency and damage assessment", in Health Monitoring of Structural and Biological Systems XIV. SPIE, 2020.
[4] H. Zheng, "Research on heat loss evaluation method of heating building envelope based on UAV infrared thermal imaging technology",

South China University of Technology, 2022.
[5] H. Cai, H. Yin, Y. Wei, et al, Establishment Method of Three-Dimensional Thermal Infrared Model of Building Based on SFM Technology, Journal of Building Energy Efficiency, 51, 10 (2023).
[6] G. Li, X. Qian, X. Qu, SOSMaskFuse: An infrared and visible image fusion architecture based on salient object segmentation mask, IEEE Transactions on Intelligent Transportation Systems, 2023.
[7] X. Li, Q. Wu, B. Xiao, et al, High-speed and robust infrared-guiding multiuser eye localization system for autostereoscopic display, Applied Optics, 59, 14 (2020).
[8] X. Li, Q. Wu, Y. Wang, Binocular vision calibration method for a long-wavelength infrared camera and a visible spectrum camera with different resolutions, Optics Express, 29, 3 (2021).
[9] J. Wu, "Research on 3D Model of Temperature Distribution Based on Thermal Infrared Image", Northeastern University, 2015.
[10] Y. Chen, "Reconstruction of infrared and optical images of buildings under UAV platform, Nanjing University of Aeronautics and Astronautics, 2019.
[11] S. Vidas, R. Lakemond, R. Denman, et al, A mask-based approach for the geometric calibration of thermal-infrared cameras, IEEE Transactions on Instrumentation and Measurement, 61, 6 (2012).
[12] R. Yang, Y. Chen, Design of a 3-D infrared imaging system using structured light, IEEE Transactions on Instrumentation and Measurement, 60, 2 (2010.
[13] Y. An, S. Zhang, High-resolution, real-time simultaneous 3D surface geometry and temperature measurement, Optics express, 24, 13 (2016.
[14] J. Rangel, S. Soldan, A. Kroll, "3D thermal imaging: Fusion of thermography and depth cameras", in International Conference on Quantitative InfraRed Thermography, 2014, 3.
[15] S. Schramm, P. Osterhold, R. Schmoll, et al, Combining modern 3D reconstruction and thermal imaging: Generation of large-scale 3D thermograms in real-time, Quantitative InfraRed Thermography Journal, 19, 5 (2022.
[16] E. Adamopoulos, M. Volinia, M. Girotto, et al, Three-dimensional thermal map** from IRT images for rapid architectural heritage NDT, Buildings, 10, 10 (2020).
[17] Y. Cao, B. Xu, Z. Ye, et al, Depth and thermal sensor fusion to enhance 3D thermographic reconstruction, Optics express, 26, 7 (2018.
[18] B. Xu, Z. Ye, F. Wang, et al, On-the-fly extrinsic calibration of multimodal sensing system for fast 3D thermographic scanning, Applied Optics, 58, 12 (2019.
[19] S. Schramm, J. Rangel, A. Kroll, "Data fusion for 3D thermal imaging using depth and stereo camera for robust self-localization", in 2018 IEEE Sensors Applications Symposium (SAS), IEEE, 2018, p. 1-6.
[20] Y. Yang, C. Xu, Fusion Reconstruction Method for 3D Temperature Fields on the Human Body Surface, Infrared Technology, 44, 1 (2022.
[21] M. M. Groz, E. Abisset-Chavanne, A. Meziane, et al, Three-dimensional reconstruction of thermal volumetric sources from surface temperature fields measured by infrared thermography, Applied Sciences, 9, 24 (2019).
[22] B. Mildenhall, P. P. Srinivasan, M. Tancik, et al, Nerf: Representing scenes as neural radiance fields for view synthesis, Communications of the ACM, 65, 1 (2021).
[23] Q. Xu, Z. Xu, J. Philip, et al, "Point-nerf: Point-based neural radiance fields", in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, p. 5438-5448.
[24] S. Fridovich-Keil, A. Yu, M. Tancik, et al, "Plenoxels: Radiance fields without neural networks", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 5501-5510.
[25] B. Kerbl, G. Kopanas, T. Leimkühler, et al, 3d gaussian splatting for real-time radiance field rendering, ACM Transactions on Graphics, 42, 4 (2023).