TeX-NeRF: Neural Radiance Fields for Novel HADAR View Synthesis

Chonghao Zhong[1], Chao Xu[1,†]

***Abstract*— Most existing NeRF methods rely on RGB images, making them unsuitable for scenarios with darkness, low light, or adverse weather conditions. To address this limitation, we propose TeX-NeRF, a NeRF framework based on heat sensing, designed for a new task: novel HADAR view synthesis. Our approach leverages Pseudo-TeX Vision to effectively transform heat sensing images through a structured mapping process. We introduce a loss function tailored to the transformed representation and incorporate temperature gradient embedding to enhance the capture of thermal information. Additionally, we construct 3D-TeX, a high-quality heat sensing dataset, to validate our method. Extensive experiments demonstrate that TeX-NeRF significantly improves pose estimation success rates for heat sensing images and outperforms existing approaches in novel HADAR view synthesis.**

I. INTRODUCTION

Since its proposal, Neural Radiance Fields (NeRF) [1] has achieved significant success in 3D reconstruction, novel view synthesis, and applications such as robotic perception, navigation [2]–[4], virtual reality, as well as computer vision tasks including semantic segmentation [5], target detection [6], etc. Most NeRF models derive implicit scene representations from RGB images captured by visible cameras. Recently, NeRF has been extended to other modalities [7]–[9]. Combining RGB with other modalities yields more accurate scene representations than using a single modality. However, heat sensing are often used as a supplementary modality to enhance RGB images, rather than being used independently. Heat sensing images face challenges such as low contrast, missing details and burring, due to sensor noise, limited pixel array size and wavelength differences between thermal and optical radiation. These challenges make camera pose estimation methods based on Structure-from-Motion (SfM) more difficult, which in turn affects the quality of subsequent novel view synthesis and 3D reconstruction. Although multimodal fusion helps mitigate some of these issues, it also increases system complexity, challenges in data synchronization and calibration, and raises computational and power costs.

To address the aforementioned challenges, we propose TeX-NeRF, a NeRF-based method that relies solely on heat sensing images, extending NeRF to the HADAR [10] modality. Heat-sensing imaging captures the emitted thermal radiation from objects. The wavelength of thermal radiation is longer than that of visible light, typically ranging from 0.7 μm to 1 mm. Visible light sensors capture reflected signals from illuminated objects, whereas heat sensing sensors detect emitted thermal radiation, revealing temperature differences independent of light sources and enabling operation in total darkness. The emitted thermal radiation is closely related to the object's surface emissivity. In ideal cases, the emissivity of a blackbody is equal to 1. However, in real-world scenarios, the surface emissivity of objects is less than 1. Most existing NeRF methods based on heat sensing aim to improve visual quality in novel view synthesis and 3D reconstruction, but they neglect the emissivity of objects and assume it to be 1, which is unreasonable. To address this issue, we refer to the NASA JPL ECOSTRESS spectral library [11] and HADAR, introducing the average surface emissivity of objects in the Long-Wave Infrared (LWIR) spectrum (i.e. 8-14 μm) as a prior. Using the emissivity e, we calculate the temperature T and texture X (as described in III). To facilitate visualization, the temperature T, emissivity e, and texture X are mapped to the HSV color space, corresponding to the saturation S, hue H, and value V, respectively. The resulting image, which we term Pseudo-TeX Vision, represents the HADAR modality and is shown

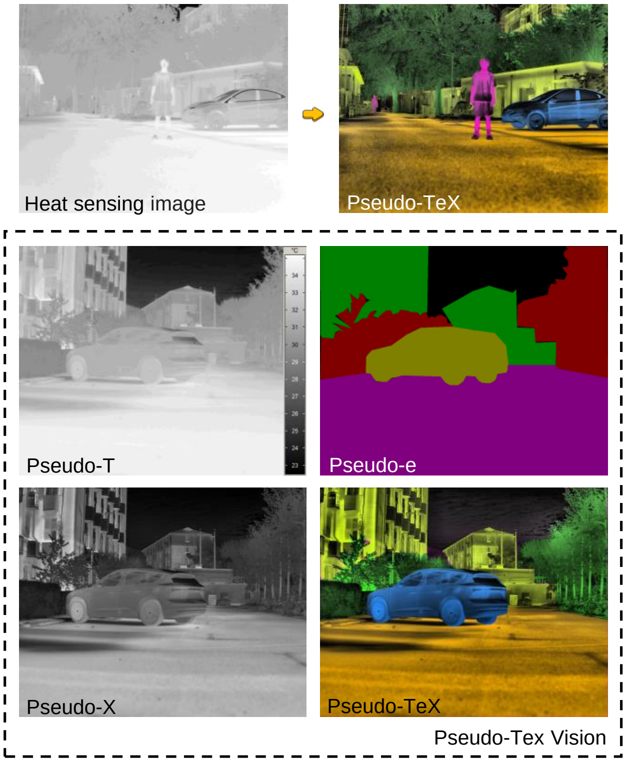

Fig. 1. Pseudo-TeX Vision Schematic: Overcoming Ghosting Effects and Recovering Texture Details in Heat Sensing

[1] Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing Institute of Technology, Beijing, China. zchnanguan7@gmail.com, rockyxu@bit.edu.cn
† Corresponding author.

in Figure 1.

Compared to original heat sensing images, Pseudo-TeX Vision images exhibit richer details and texture information, which also makes them more suitable for observation. Experimental results demonstrate that Pseudo-TeX Vision effectively addresses the issues in original heat sensing images, such as low contrast, blurred details, and missing textures, which prevent the use of COLMAP [12] for camera pose estimation. This improvement provides accurate camera poses for subsequent processing. Furthermore, our TeX-NeRF, which is built upon Nerfacto in the Nerfstudio framework [13], is designed for novel HADAR view synthesis. Extensive experiments conducted on both our self-collected 3D-TeX dataset and public datasets show that TeX-NeRF effectively adapts to Pseudo-TeX Vision, significantly enhancing the quality and accuracy of novel HADAR view synthesis. Ablation studies further validate the effectiveness of the proposed improvements. In summary, the main contributions of this work are as follows:

- We propose TeX-NeRF, a Neural Radiance Field (NeRF) designed for novel HADAR view synthesis, which solely relies on heat sensing, incorporates surface emissivity information, and employs Pseudo-TeX Vision to process heat sensing images.
- We introduce the 3D-TeX dataset, consisting of 1000 pairs of high-quality heat sensing images and their corresponding Pseudo-TeX Vision images.
- Extensive experimental results validate the feasibility and effectiveness of the proposed method.

II. RELATED WORK

A. Heat Sensing NeRF and 3DGS

NeRF shows great potential in the synthesis and transformation of cross-modal data. In recent studies, NeRF has been extended to combine data from other modalities (e.g. depth maps, heating sensing images, etc.) with RGB images to generate richer 3D representations [7] [14]. ThermoNeRF [15] uses paired RGB and heat sensing images to learn the scene density, while using different networks to estimate color and temperature information to overcome the lack of texture in thermal images. Thermal NeRF [16] proposes structural heat constraints acting on heat distributions, leveraging heat sensing features for high-fidelity 3D representation.

3D Gaussian Splatting [17] enhances NeRF by representing each point in 3D space as a Gaussian distribution with position, color, and material properties, making scene representation more compact, efficient, and capable of generating high-resolution views with improved rendering speed and quality. Thermal3D-GS [18] addresses the unique physical properties of thermal radiation for novel view synthesis, overcoming issues such as floating artifacts and blurred edges in the 3D Gaussian Splatting process. ThermalGS [19] tackles the challenge of dynamic thermal 3D reconstruction by using 3D Gaussian Splatting, capturing both spatial and temporal variations in thermal radiation.

B. TeX Vision

TeX Vision overcomes the ghosting effect, improves object detection and ranging capabilities and provides additional physical attributes [20]. TeX Vision, first proposed in HADAR [10], uses a hyperspectral camera in the LWIR spectrum to obtain the object's emissivity, enabling material identification and reflective feature extraction. TeX Vision utilizes hyperspectral thermal cubes to estimate the material classes of objects, subsequently assigning emissivity information. NeRF2Physics [21] and PUGS [22] estimate object categories using NeRF and 3DGS methods, respectively, but both rely on visible light images as input.

C. Semantic Segmentation for heat sensing Images

EC-CNN [23] leverages edge prior information to enhance segmentation quality. Wang et al. [24] proposed a heat sensing pedestrian segmentation algorithm including a conditional generative adversarial network. Nightvision-Net (NvNet) [25] introduces data refinement and data normalization to improve the segmentation performance. Multi-level correction network (MCNet) [26] proposes a multi-level attention module (MAM) to solve the problem of low resolution and blurred edges. Feature Transverse Network (FTNet) [27] introduced an end-to-end trainable convolutional neural network for reliable pixel-level classification. UNIP [28] benchmarks heat sensing segmentation performance across pre-training strategies and proposes an effective selective knowledge distillation approach for domain transfer. Additionally, foundational models like SAM [29] and SAM2 [30] have also been effectively applied in the field of heat sensing segmentation [31]–[33].

III. METHOD

A. Overview

As shown in Figure 2, our proposed method takes heat sensing images as input to obtain the 3D scene representation of Pseudo-TeX Vision. The method can be divided into two stages: the generation of Pseudo-TeX Vision images (Section III-B) and the training of TeX-NeRF (Section III-C). First, we obtain the heat sensing image masks using SAM2 [30], fine-tuned on heat sensing dataset. We then use CLIP [34] to generate object category labels corresponding to the masks. By querying a lookup table, constrcted based on the NASA JPL ECOSTRESS spectral library, we retrieve the average surface emissivity in the LWIR spectrum for each object category. This information is used to compute temperature and texture data. The results are processed throught a mapping procedure to generate Pseudo-TeX Vision images, which are then used to obtain camera poses through the SfM algorithm. The encoded camera poses are concatenated with the temperature gradient embedding and input into the TeX-NeRF network for training.

B. Pseudo-TeX Image Generation

According to Planck's law, all natural and man-made objects emit thermal radiation at a given temperature, with the

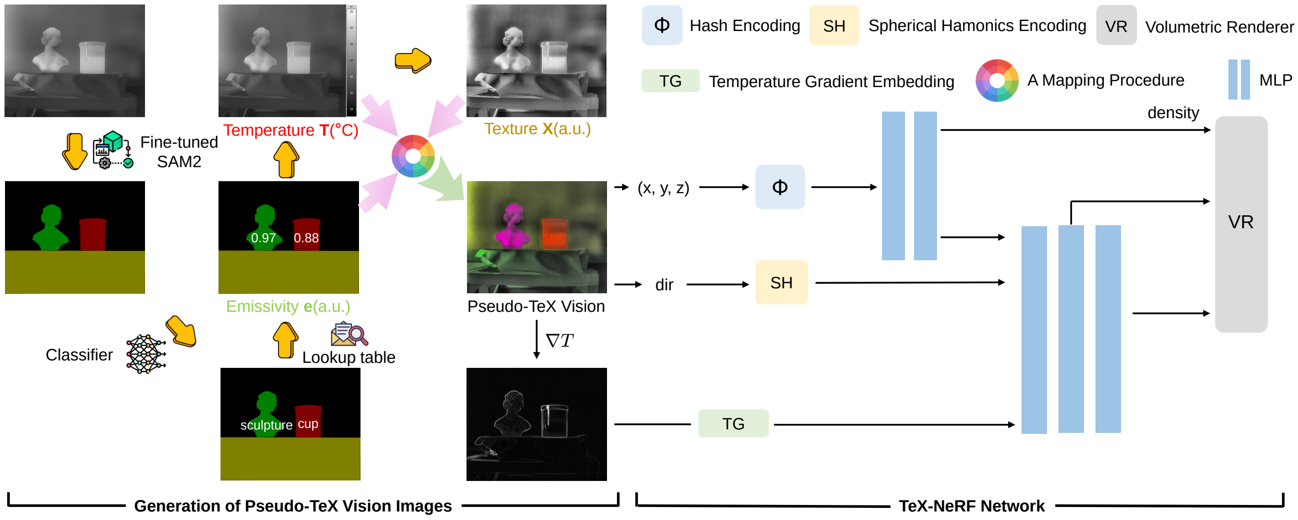

Fig. 2. **Overview of the proposed method.** The pipeline consists of two main stages: (1) generation of Pseudo-TeX Vision images and (2) training of the TeX-NeRF network. Heat sensing images are first processed using a fine-tuned SAM2 model to obtain object masks, followed by the use of CLIP to generate object category labels. Material properties and average surface emissivity are then retrieved from a lookup table based on object categories, enabling the computation of temperature and texture data. These results are processed through a mapping procedure to create Pseudo-TeX Vision images, which are then used as input for training TeX-NeRF to achieve 3D scene representation. The camera poses are encoded, and the result is concatenated with the temperature gradient embedding of temperature T from the Pseudo-TeX Vision images before being fed into TeX-NeRF for training

intensity depending on the object's emissivity. The thermal radiance equation is given in Eq.1.

$$S_{\alpha\nu} = e_{\alpha\nu} B_\nu(T_\alpha) + [1 - e_{\alpha\nu}] X_{\alpha\nu}, \tag{1}$$

Where $S_{\alpha\nu}$ represents the thermal radiation signal of the object α captured by the heat sensing sensor at wave number ν, and e represents the emissivity of the object, which is highly correlated with the material of the object. T_α represents the temperature of the object. B_ν is Planck's formula for blackbody radiation, see Eq.2. $X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta} S_{\beta\nu}$ The $V_{\alpha\beta}$ is the thermal illumination factor, which reflects what percentage of the thermal radiation signal reflected onto an object can be collected by the sensor. Since there are many objects in the environment, it is reflected in the equation as cumulative, representing the total amount of thermal radiation signal reflected off the objects in the environment.

$$B(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{\frac{h\nu}{ek_B T} - 1} \tag{2}$$

Notice that the thermal radiation signal $S_{\alpha\nu}$ can be expressed in terms of three physical quantities: temperature T, emissivity e, and texture X. However, as discussed in HADAR [10], this process introduces what is known as TeX degeneracy, where different combinations of T, e, and X can produce the same thermal radiance signal S. This makes solving for T, e, and X individually from S a highly ill-posed inverse problem with potentially infinite solutions.

Solving ill-posed inverse problems often requires prior knowledge. In real-world scenarios, most objects are man-made and adhere to consistent industrial standards. This means that their emissivity is highly consistent when they are made of the same material. By performing semantic segmentation on heat sensing images, we obtain the object category information and retrieve the corresponding average surface emissivity in the LWIR band from the NASA JPL ECOSTRESS spectral library, as shown in Figure 3. Thus, the surface emissivity in the LWIR band is used as prior knowledge for the target scene objects.

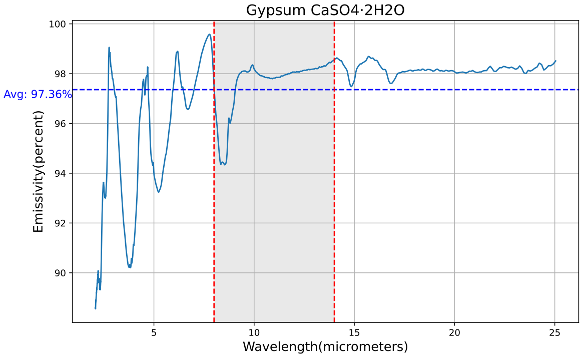

Fig. 3. Example from the NASA JPL ECOSTRESS spectral library [11]. Average surface emissivity in the LWIR band for gypsum

However, directly utilizing the pretrained SAM2 model [30] for heat sensing semantic segmentation does not yield satisfactory performance due to the significant domain gap between visible and heat sensing images. To address this, we fine-tune the Sam2-hiera-s model using self-collected 3D-TeX, SODA [35], and SCUT-Seg [36] datasets, enabling the generation of accurate masks for heat sensing images. Once the mask is obtained, we use CLIP [34] to classify the object within the scene's category range. Based on the classified category, we retrieve the corresponding average surface emissivity for the object in the LWIR spectrum from the spectral library (e.g., assigning sculptures to gypsum and cups to plastic).

Once e is known, it follows from Eq.3 that only S and e will survive when we take the derivative of the wavelength

ν:

$$[S/(1-e)]^{'} = [e/(1-e)]^{'} B_\nu(T) \quad (3)$$

here, the subscript is omitted in Eq.3 to aviod confusion, where prime indicates the derivative with respect to wavenumber ν. Since S is observable and e is known, the temperature T can be obtained by solving Eq.2. Transforming Eq.1 gives:

$$V_0 = \frac{S_\nu - e_\nu B_\nu(T)}{(1-e_\nu)B_\nu(T_0)} \quad (4)$$

When e and T are known, V can be determined from Eq.4. Finally, X is calculated using $X = \int V_0 B_\nu(T_0)\mathrm{d}\nu$. Thus, the three physical quantities T, e and X are obtained.

After obtaining the temperature T, emissivity e, and texture X, they are mapped into the HSV color space. In contrast to the RGB color space, where each channel represents light intensity, each channel in the HSV space used in Pseudo-TeX Vision has a physical meaning. Specifically, the Hue channel is strongly correlated with the semantic information of an object. For example, blue typically represents water or the sky, green represents grass or leaves, and red indicates heat-generating objects. To reconstruct or simulate visible RGB images using thermal radiance signals, in Pseudo-TeX Vision, the object's material category m is represented in the Hue channel (which can also be viewed as e due to their strong correlation), the temperature T is displayed in the saturation channel, and the texture X is shown in the value channel.

Based on the category labels of the masks output by CLIP, the label not only determines the value of the Hue (H) channel within the same category mask but also is used to look up the corresponding emissivity. Once the emissivity is known, the temperature T and texture X are computes as described above. The values of T, e and X are then mapped to the saturation (S), hue (H) and value (V) channels, respectively, to generate the Pseudo-TeX Vision images.

C. TeX-NeRF for Novel HADAR View Systhesis

Preliminary of Neural Radiance Field. The core principle of Neural Radiance Fields (NeRF) is to model a 3D scene's radiance field using an MLP to generate realistic novel views, levering volume rendering for synthesizing density and color, combined with ray tracing techniques.

In predicting volume density and color, NeRF's MLP takes the 3D spatial coordinates $\mathbf{x}$ and the viewing direction $\mathbf{d}$ of each sampled point on the ray, and outputs the corresponding volume density and color. (See Eq.5.)

$$(\sigma, \mathbf{c}) = \mathrm{MLP}(\mathbf{x}, \mathbf{d}) \quad (5)$$

The σ represents the point's density, and $\mathbf{c}$ represents the point's RGB color value. The final color $c(r)$ of each ray is computed by accumulating multiple sampled points along the ray. For a ray $r(t)$, the final color is calculated using the following volume rendering equation:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1-\exp(-\sigma_i\delta_i))\mathbf{c}_i \quad (6)$$

The term $T_i = \exp\left(-\sum_{j=1}^{i-1}\sigma_j\delta_j\right)$ denotes the transmittance of light from the first point to the i-th point, representing the probability that the light is not blocked by previous points. Here, σ_i is the volume density at the i-th sampling point, $\mathbf{c}_i$ is the color of the i-th sample, and δ_i is the distance between neighboring sampling points.

TeX-NeRF Rendering in Pseudo-TeX Vision. In the original NeRF, the output of the MLP consists of the volume density σ and the color $c = (r, g, b)$. Since Pseudo-TeX vision maps heat sensing images to HSV color space, by changing the RGB representation to HSV representation, we expect the color branch output of the MLP to be (H, S, V), with the volume density remaining unchanged:

$$(\sigma, H, S, V) = MLP(x, d) \quad (7)$$

where H stands for hue, which indicates the type of color and is typically in the [0°, 360°] range. S stands for saturation, which represents the intensity of the color and is in the [0,1] range, and V stands for value, which indicates the brightness of the color, and is also in the [0,1] range. During volume rendering, $c_i = (H_i, S_i, V_i)$, and the rendering formula calculates the color contribution of each point from the HSV value:

$$C_{HSV}(r) = \sum_{i=1}^{N} T_i(1-exp(-\sigma_i\delta_i))(H_i, S_i, V_i) \quad (8)$$

where (H_i, S_i, V_i) are the hue, saturation and values of the point.

TeX-NeRF is an improved version of the Nerfacto method in Nerfstudio, as shown in Figure 2. It outputs the HSV values for each pixel by encoding position and direction information using hash encoding and spherical harmonics encoding. This approach enhances the model's ability to express spatial position and directional changes, improving both efficiency and accuracy during the rendering and generation processes. The hue (H) channel represents the type of color and exhibits periodic and smooth changes. Given that the variations in H are generally structured and less complex conpared to other color representations, we use a relatively small network complexity to model these variations efficiently. TeX-NeRF achieves good performance by outputting the hue value in the second layer of the MLP. Since the values of H, S, and V are normalized to the [0,1] range, we apply a sigmoid activation function in the output layer to constrain the network's predictions to this interval.

When the output space is changed to HSV, the loss function needs to account for this change. Since hue is a cyclic space, it typically takes values in the range [0, 1]. (While hue itself is periodic, the assignment of specific values to the hue channel based on object categories is simply for modeling purposes and does not imply that the object categories themselves are cyclic.) Therefore, the color phase loss must consider the periodicity of hue, as directly computing the difference between two values may overlook cases where the values span across 0 and 1. To address this issue, the following loss function is used:

$$\mathcal{L}_H = \mathrm{SoftMin}\left(\left|\mathbf{H}^{\mathrm{pred}} - \mathbf{H}^{\mathrm{gt}}\right|, 1 - \left|\mathbf{H}^{\mathrm{pred}} - \mathbf{H}^{\mathrm{gt}}\right|\right) \tag{9}$$

The SoftMin function is defined as:

$$\mathrm{SoftMin}(a, b) = -\frac{1}{\beta} \log\left(e^{-\beta a} + e^{-\beta b}\right) \tag{10}$$

where β is a parameter that controls the smoothness of the transition. In our experiments, β is chosen as 10 to smooth the hue loss and ensure continuity and differentiability at the periodic boundaries.

Saturation and value do not suffer from the above problem, the loss can be calculated using the standard MSE. So, the final loss function can be written as:

$$\mathcal{L}_{HSV} = \frac{1}{N} \sum_{i=1}^{N} \Big(\mathcal{L}_H + \parallel \mathbf{S}_i^{\mathbf{pred}} - \mathbf{S}_i^{\mathbf{gt}} \parallel^2$$

$$+ \parallel \mathbf{V}_i^{\mathbf{pred}} - \mathbf{V}_i^{\mathbf{gt}} \parallel^2\Big) \tag{11}$$

Temperature Gradient Embedding. In our method, the temperature gradient is computed by applying the Laplacian operator (see Eq.12) to the S-channel of the Pseudo-TeX Vision image, resulting in a $h \times w$ grayscale embedding. Since the model's input requires a tensor of shape $batchsize \times c$. We use a two-layer MLP to map the gradient embedding to c-dimensional space, which is then concatenated with other encoded features along the feature dimension.

$$\nabla T = \nabla^2 S \tag{12}$$

where ∇^2 represents the Laplacian operator. The resulting embedding is then combined with pose and direction features to form the final input tensor.

IV. EXPERIMENTS

This section begins with an introduction to the datasets used in the experiments. We then present the camera pose estimation experiment, which demonstrates the advantage of Pseudo-TeX Vision images over original heat sensing images. Qualitative and quantitative comparisons further validate the effectiveness of TeX-NeRF in reconstructing Pseudo-TeX Vision images. Finally, ablation studies highlight the improvements introduced by TeX-NeRF.

A. Datasets

3D-TeX Dataset. Due to the lack of high-quality 3D reconstruction datasets for heat sensing images, we have created the 3D-TeX dataset. The data was collected using the iRAY LGCS121 sensor, which has a resolution of 1280x1024 pixels and operates within a wavelength range of 8 to 14 µm. The 3D-TeX dataset consists of 1000 pairs of heat sensing images and their corresponding Pseudo-TeX Vision images, along with pixel-wise mask annotations for the heat sensing images. The dataset includes three distinct objects: a gypsum sculpture, a plastic container half-filled with hot water, and a heating table maintaining a constant temperature of 55°C . Data was captured under both illuminated and dark environments, with additional temperature variation introduced using a heater. This dataset will be made publicly available to support further research and applications.

ThermalMix Dataset. ThermalMix [16] is a multi-view, object-centered dataset containing RGB and thermal images of six common objects. For this experiment, we select heat sensing images of three scenes (hands, pans, and laptops) to validate the method. The hand dataset contains 42 images (22.1 to 33.6°C), the pan dataset includes 82 images (19.4 to 129.8°C), and the laptop dataset contains 93 images (21.6 to 40.2°C).

B. Camera Pose Estimation

We applied the COLMAP (version 3.11) [12] to both heat sensing images and images processed through Pseudo-TeX Vision for camera pose estimation, as shown in the Table I. In four scene: heating table, sculpture with a cup, laptop and hand, the heat sensing images, characterized by low contrast and limited texture information, resulted in a relatively low success rate for pose estimation. In contrast, images processed by Pseudo-TeX Vision exhibited a significant improvement in pose estimation accuracy. This demonstrates that our method not only preserves the information from the original heat sensing images but also enhances details, texture, and contrast, making it highly beneficial for 3D scene reconstruction based on heat sensing images.

TABLE I

CAMERA POSE ESTIMATION RESULTS FOR ORIGINAL HEAT SENSING AND PSEUDO-TEX VISION PROCESSED IMAGES

scene	Heat Sensing	Pseudo-TeX Vision
heating table	7/183	183/183
scu & cup	2/61	61/61
laptop	4/93	91/93
hand	3/42	42/42

C. Novel HADAR View Synthesis with TeX-NeRF

We conducted novel HADAR view synthesis experiments using TeX-NeRF on the 3D-TeX and ThermalMix datasets. For the 3D-TeX dataset, we selected two scenes: Scene 1, which contains a constant temperature heating table, and Scene 2, which includes a sculpture and a container. We reserved 10% of the total images in each scene as the test set. The experiments, based on Nerfstudio [13], compare TeX-NeRF with advanced RGB-based methods, including Instant-NGP [37], Mip-NeRF [38], Nerfacto, and Splatfacto. These methods first convert the Pseudo-TeX Vision images from HSV to RGB before proceeding with the subsequent training and testing processes. Qualitative results are shown in Figure 4. Additionally, we evaluated the novel view synthesis quality using quantitative metrics such as PSNR, SSIM, and LPIPS, as shown in Table II. Since the variations in hue (H) and saturation (S) do not directly affect the structure, but can still influence the perceptual quality of the image, we computed the SSIM only on the value (V) channel. The results demonstrate that TeX-NeRF significantly outperforms other methods in both visual quality and evaluation metrics, especially in novel HADAR view synthesis.

TABLE II

Quantitative Evaluation Results on Different Scenes

Metric	Method	Scene 1	Scene 2	hand	laptop	pan	Avg
PSNR↑	Instant-ngp	21.01	21.69	16.99	14.00	18.21	18.38
	Mip-nerf	20.74	19.88	16.31	17.62	19.59	18.83
	Nerfacto	18.39	19.43	14.42	16.46	20.18	17.78
	Splatfacto	21.90	24.05	18.36	**21**	23.87	21.84
	TeX-NeRF(ours)	**24.97**$_{+6.58}$	**32.72**$_{+13.29}$	**20.01**$_{+5.59}$	20.91$_{+4.45}$	**27.21**$_{+7.03}$	**25.16**$_{+7.38}$
SSIM↑	Instant-ngp	0.70	0.55	0.58	0.48	0.64	0.59
	Mip-nerf	0.66	0.54	0.60	0.53	0.67	0.60
	Nerfacto	0.65	0.54	0.51	0.52	0.62	0.57
	Splatfacto	0.78	0.53	0.55	**0.65**	**0.84**	0.67
	TeX-NeRF(ours)	**0.83**$_{+0.18}$	**0.88**$_{+0.34}$	**0.65**$_{+0.14}$	0.62$_{+0.10}$	0.72$_{+0.10}$	**0.74**$_{+0.17}$
LPIPS↓	Instant-ngp	0.69	0.81	0.62	0.81	0.44	0.67
	Mip-nerf	0.67	0.61	0.43	0.56	0.41	0.54
	Nerfacto	0.66	0.59	0.46	0.52	0.42	0.53
	Splatfacto	0.51	0.54	0.30	**0.37**	**0.27**	0.40
	TeX-NeRF(ours)	**0.33**$_{-0.33}$	**0.24**$_{-0.35}$	**0.23**$_{-0.23}$	0.39$_{-0.13}$	0.32$_{-0.10}$	**0.30**$_{-0.23}$

The red subscript values indicate the performance gains of TeX-NeRF over Nerfacto, as our method is built upon Nerfacto.

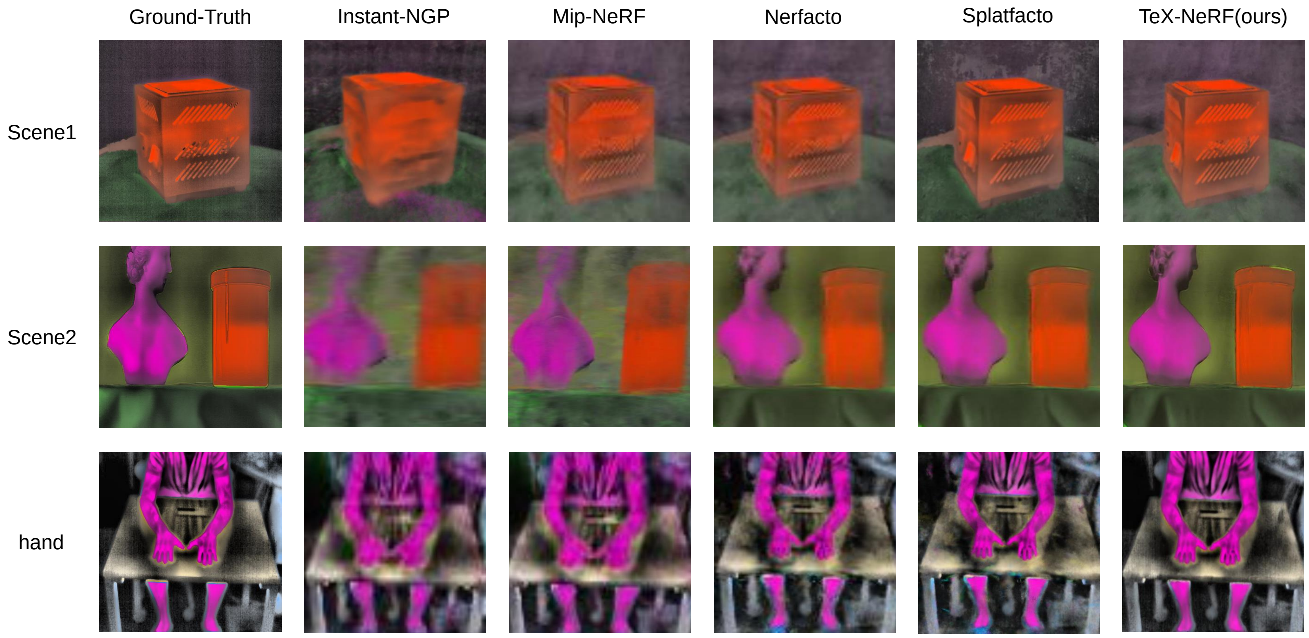

Fig. 4. Qualitative Comparison of TeX-NeRF with RGB-based Methods on 3D-TeX Dataset

D. Ablation Study of TeX-NeRF

We validate the design choices of our algorithm through an ablation study, as shown in Table III. The results are presented for scene 1 of our 3D-TeX dataset. *Nerfacto RGB* represents the vanilla Nerfacto, which renders images using the RGB color space. *Nerfacto w/* $\mathcal{L}_{HSV}$ indicates the use of TeX-NeRF's loss function in Nerfacto. *TeX-NeRF w/o TG* means the model without the temperature gradient embedding. *Nerfacto* H_3 refers to the case where the hue channel is directly output from the final layer of the MLP, and *TeX-NeRF* refers to our proposed method. The ablation study results show that TeX-NeRF significantly outperforms in terms of reconstruction quality in the HSV color space, and the various improvements effectively enhance the model's understanding of Pseudo-TeX Vision images.

TABLE III

Ablation Study Results on TeX-NeRF Design Choices

Method	PSNR↑	SSIM↑	LPIPS↓
Nerfacto RGB	18.39	0.65	0.66
Nerfacto w/ $\mathcal{L}_{HSV}$	20.81	0.69	0.63
TeX-NeRF w/o TG	22.74	0.73	0.59
TeX-NeRF H_3	22.05	0.82	0.41
TeX-NeRF(ours)	24.97	0.83	0.33

V. Conclusion

In this paper, we introduce TeX-NeRF, a novel NeRF-based approach that operates exclusively on heat sensing images. We use Pseudo-TeX Vision to process the heat sensing images, which facilitates effective scene representation. Extensive experiments and visualizations further demonstrate this processing significantly improves the success rate of camera pose estimation. Moreover, TeX-NeRF exhibits a distinct advantage in representing scenes within HADAR modality, which is particularly beneficial for applications such as robotic perception in low-light or dark environments. In future work, we aim to extend the method's robustness and accuracy across a wider variety of object categories, further broadening its applicability. Additionally, we plan to explore the potential of Gaussian Splatting for HADAR, investigating its capability to enhance efficiency and rendering quality in our framework.

References

[1] 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.

[2] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 786–12 796.

[3] Z. Zhu, Y. Chen, Z. Wu, C. Hou, Y. Shi, C. Li, P. Li, H. Zhao, and G. Zhou, "Latitude: Robotic global localization with truncated dynamic low-pass filter in city-scale nerf," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 8326–8332.

[4] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, "Vision-only robot navigation in a neural radiance world," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4606–4613, 2022.

[5] F. Liu, C. Zhang, Y. Zheng, and Y. Duan, "Semantic ray: Learning a generalizable semantic field with cross-reprojection attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 386–17 396.

[6] B. Hu, J. Huang, Y. Liu, Y.-W. Tai, and C.-K. Tang, "Nerf-rpn: A general framework for object detection in nerfs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 528–23 538.

[7] M. Özer, M. Weiherer, M. Hundhausen, and B. Egger, "Exploring multi-modal neural scene representations with applications on thermal imaging," *arXiv preprint arXiv:2403.11865*, 2024.

[8] M. Poggi, P. Z. Ramirez, F. Tosi, S. Salti, S. Mattoccia, and L. Di Stefano, "Cross-spectral neural radiance fields," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 606–616.

[9] J. Zhang, F. Zhang, S. Kuang, and L. Zhang, "Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7178–7186.

[10] F. Bao, X. Wang, S. H. Sureshbabu, G. Sreekumar, L. Yang, V. Aggarwal, V. N. Boddeti, and Z. Jacob, "Heat-assisted detection and ranging," *Nature*, vol. 619, no. 7971, pp. 743–748, 2023.

[11] A. M. Baldridge, S. J. Hook, C. Grove, and G. Rivera, "The aster spectral library version 2.0," *Remote sensing of environment*, vol. 113, no. 4, pp. 711–715, 2009.

[12] 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.

[13] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.

[14] H. Zhu, Y. Sun, C. Liu, L. Xia, J. Luo, N. Qiao, R. Nevatia, and C.-H. Kuo, "Multimodal neural radiance field," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9393–9399.

[15] M. Hassan, F. Forest, O. Fink, and M. Mielle, "Thermonerf: Multimodal neural radiance fields for thermal novel view synthesis," *arXiv preprint arXiv:2403.12154*, 2024.

[16] T. Ye, Q. Wu, J. Deng, G. Liu, L. Liu, S. Xia, L. Pang, W. Yu, and L. Pei, "Thermal-nerf: Neural radiance fields from an infrared camera," *arXiv preprint arXiv:2403.10340*, 2024.

[17] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[18] Q. Chen, S. Shu, and X. Bai, "Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 253–269.

[19] Y. Liu, X. Chen, S. Yan, Z. Cui, H. Xiao, Y. Liu, and M. Zhang, "Thermalgs: Dynamic 3d thermal reconstruction with gaussian splatting," *Remote Sensing*, vol. 17, no. 2, p. 335, 2025.

[20] F. Bao, S. Jape, A. Schramka, J. Wang, T. E. McGraw, and Z. Jacob, "Why thermal images are blurry," *Optics Express*, vol. 32, no. 3, pp. 3852–3865, 2024.

[21] A. J. Zhai, Y. Shen, E. Y. Chen, G. X. Wang, X. Wang, S. Wang, K. Guan, and S. Wang, "Physical property understanding from language-embedded feature fields," in *CVPR*, 2024.

[22] Y. Shuai, R. Yu, Y. Chen, Z. Jiang, X. Song, N. Wang, J. Zheng, J. Ma, M. Yang, Z. Wang, *et al.*, "Pugs: Zero-shot physical understanding with gaussian splatting," *arXiv preprint arXiv:2502.12231*, 2025.

[23] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3069–3082, 2020.

[24] P. Wang and X. Bai, "Thermal infrared pedestrian segmentation based on conditional gan," *IEEE transactions on image processing*, vol. 28, no. 12, pp. 6007–6021, 2019.

[25] S. Chen, Z. Chen, X. Xu, N. Yang, and X. He, "Nv-net: Efficient infrared image segmentation with convolutional neural networks in the low illumination environment," *Infrared Physics & Technology*, vol. 105, p. 103184, 2020.

[26] H. Xiong, W. Cai, and Q. Liu, "Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," *Infrared Physics & Technology*, vol. 113, p. 103628, 2021.

[27] K. Panetta, K. S. Kamath, S. Rajeev, and S. S. Agaian, "Ftnet: Feature transverse network for thermal image semantic segmentation," *IEEE Access*, vol. 9, pp. 145 212–145 227, 2021.

[28] T. Zhang, J. Wen, Z. Chen, K. Ding, S. Xiang, and C. Pan, "Unip: Rethinking pre-trained attention patterns for infrared semantic segmentation," *arXiv preprint arXiv:2502.02257*, 2025.

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[30] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[31] M. Zhang, Y. Wang, J. Guo, Y. Li, X. Gao, and J. Zhang, "Irsam: Advancing segment anything model for infrared small target detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 233–249.

[32] Y. F. A. Gaus, N. Bhowmik, B. K. Isaac-Medina, and T. P. Breckon, "Performance evaluation of segment anything model with variational prompting for application to non-visible spectrum imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3142–3152.

[33] X. Yang, H. Dai, Z. Wu, R. Bist, S. Subedi, J. Sun, G. Lu, C. Li, T. Liu, and L. Chai, "Sam for poultry science," *arXiv preprint arXiv:2305.10254*, 2023.

[34] 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.

[35] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3069–3082, 2020.

[36] H. Xiong, W. Cai, and Q. Liu, "Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," *Infrared Physics & Technology*, vol. 113, p. 103628, 2021.

[37] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[38] 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.