Exploring Multi-modal Neural Scene Representations With Applications on Thermal Imaging

Mert Özer, Maximilian Weiherer, Martin Hundhausen, and Bernhard Egger

Friedrich-Alexander-Universität Erlangen-Nürnberg
firstname.lastname@fau.de

Abstract. Neural Radiance Fields (NeRFs) quickly evolved as the new de-facto standard for the task of novel view synthesis when trained on a set of RGB images. In this paper, we conduct a comprehensive evaluation of neural scene representations, such as NeRFs, in the context of multi-modal learning. Specifically, we present four different strategies of how to incorporate a second modality, other than RGB, into NeRFs: (1) training from scratch independently on both modalities; (2) pre-training on RGB and fine-tuning on the second modality; (3) adding a second branch; and (4) adding a separate component to predict (color) values of the additional modality. We chose thermal imaging as second modality since it strongly differs from RGB in terms of radiosity, making it challenging to integrate into neural scene representations. For the evaluation of the proposed strategies, we captured a new publicly available multi-view dataset, *ThermalMix*, consisting of six common objects and about 360 RGB and thermal images in total. We employ cross-modality calibration prior to data capturing, leading to high-quality alignments between RGB and thermal images. Our findings reveal that adding a second branch to NeRF performs best for novel view synthesis on thermal images while also yielding compelling results on RGB. Finally, we also show that our analysis generalizes to other modalities, including near-infrared images and depth maps. Project page: https://mert-o.github.io/ThermalNeRF/.

Keywords: Multi-modal Learning · NeRF · Thermal Imaging

1 Introduction

Novel view synthesis pertains to the generation of new perspectives from an existing set of images. Historically, this problem has been tackled using conventional techniques like structure-from-motion [32], multi-view stereo [48], or image-based rendering techniques [35], and, recently, through the adoption of neural networks, prominently Neural Radiance Fields (NeRFs) [23]. NeRFs offer a paradigm shift by encapsulating the scene within a continuous radiance field, allowing the representation of volume density and view-dependent RGB colors in a four-dimensional space.

On the other hand, multi-modal imaging, characterized by the simultaneous acquisition and processing of multiple data types from different *optical sensors*

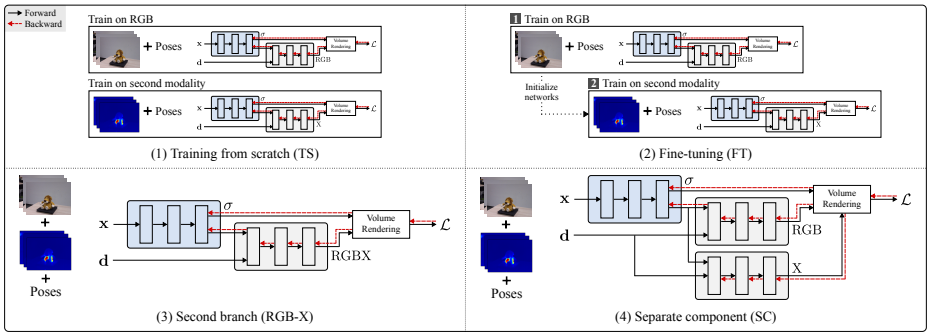

Fig. 1: Overview of the four strategies that we compare within this work. In the first strategy (TS), we train a NeRF-like base model (Instant-NGP [24] in our case) from scratch, separately for RGB and the second modality. In the second strategy (FT), we first pre-train our base model on RGB data and then fine-tune on images of the second modality. While RGB-X adds a second branch, strategy four (SC) adds an extra *network* to predict values of the additional modality. Note that RGB-X and SC yield a *single*, multi-modal scene representation, whereas TS and FT always result in two separate models, one for each modality.

(note that this definition explicitly excludes, *e.g.*, text) has shown its significance across myriad applications, ranging from surface reconstruction [4, 12, 51] and image segmentation [6] to applications in remote sensing [18, 44] and medical imaging [25, 41]. We believe that the integration of multi-modal information into modern neural scene representations like NeRFs could potentially enhance the depth and richness of the synthesized views, offering more detailed and nuanced scene reconstructions. Indeed, this has been confirmed by a recent line of work, attempting to build multi-modal NeRFs from RGB and depth information [5, 11, 13, 22, 28, 40, 50] as well as RGB and near-infrared images [11, 27].

Through recent advancements, NeRFs have undergone transformative improvements [9]. However, transitioning into a multi-modal environment introduces a host of complexities. Notably, camera pose estimation for non-RGB images becomes a formidable challenge. Established methods such as structure-from-motion can falter due to the unique features intrinsic to multi-modal (especially, multi-spectral) images. Furthermore, the alignment of multi-sensory imagery necessitates a representation in a common coordinate system, requiring either offline cross-modality calibration or *learning* of relative transformations between multiple sensors during training. Ultimately, it is also unclear how to best integrate multi-modality into modern neural scene representations; a question, which is, to date, largely unexplored and that we will address in this work.

In this paper, we conduct a comprehensive evaluation of neural scene representations, specifically, NeRFs, within a multi-modal context. We propose four different strategies of how to include a second modality (other than RGB) into a NeRF-like scene representation: (1) training from scratch, (2) fine-tuning, (3) adding a second branch, and (4) adding a separate component, see Fig. 1. We

chose thermal (*i.e.*, far-infrared) imaging as the second modality for this work, since we consider modeling thermal images, among all existing imaging modalities, to be one of the hardest. Compared to RGB images, thermal images are extremely feature-less (also, most of the features are blurry), and exhibit relatively low texture resolution even when captured with high-end cameras. As we demonstrate in the supp. material, this causes serious problems in estimating (reliable) camera poses. Moreover, because thermal images look so different compared to RGB, the underlying scene's geometry will differ from an RGB-derived geometry, rendering it non-trivial to design a neural scene representation that combines both modalities. We evaluate the proposed strategies on a newly captured, object-centric dataset that includes multi-view RGB and thermal images of six common objects. Our dataset, which we name *ThermalMix* (because it is a mix of RGB and thermal images), consists of three forward-facing and three 360-degree scenes and comes with approximately 360 images in total. As opposed to recent related works [11, 27], we employ cross-modality calibration, yielding almost perfectly aligned RGB and thermal images. Finally, we show that our results also generalize to other modalities, including near-infrared images, ultimately covering the whole bandwidth of the infrared spectrum within this work.

From a practical perspective, there is a huge number of potential domains and applications for which multi-modal neural scene representations integrating RGB *and* thermal imagery could be of interest. For instance, thermal imaging was successfully employed in agriculture [8,16,19,42], medicine [1,31,33,36], plant sciences [17, 26, 37], aviation [52], defense systems [2], food industry [10], and more (see also surveys [7, 29, 46]). Recently, a combination of RGB and thermal images was used for semantic segmentation [15, 21, 34], defect detection [47], traffic monitoring [3], 3D reconstruction in medicine [36], and food segmentation [30]. Virtually all of the aforementioned applications (in which a combination of RGB and thermal data has proven to be beneficial) can be transferred into 3D, thus allowing for, *e.g.*, better food segmentation in 3D scenes.

In summary, the core contributions of this paper are three-fold:

- We present a comprehensive study comparing four different strategies of how to learn multi-modal NeRFs based on RGB and thermal imagery.
- We propose the first *multi-view* dataset, named *ThermalMix*, of high-quality aligned RGB and thermal images captured from six common objects.
- We demonstrate that our results also generalize to other modalities, including near-infrared images and depth maps.

Our dataset is publicly available to foster future research and to serve as a benchmark, see https://mert-o.github.io/ThermalNeRF/.

2 Related Work

Neural Radiance Fields (NeRFs). NeRF [23] utilizes a continuous function F to characterize a scene, built upon simple, fully connected neural network layers. Given a spatial coordinate $\mathbf{x} \in \mathbb{R}^3$ and its associated viewing direction

$\mathbf{d} \in \mathbb{S}^2$, NeRF deduces both, the radiance $\mathbf{c}$ and density σ as $F(\mathbf{x}, \mathbf{d}; \Theta) = (\mathbf{c}, \sigma)$, where Θ stands for the parameters of the neural network. For determining the final color of a pixel, ray marching is central. Specifically, the accumulated color $\hat{\mathbf{c}}(\mathbf{r})$ of a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is computed using volume rendering:

$$\hat{\mathbf{c}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \mathbf{c}_i, \quad \text{where} \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \tag{1}$$

and $\delta_i = t_{i+1} - t_i$. Capturing high-frequency spatial variations in scenes hinges on positional encoding of spatial locations, $\mathbf{x}$ (and also viewing directions, $\mathbf{d}$). The positional encoding introduced in NeRF relies on sinusoidal oscillations to scale frequencies logarithmically, thus enabling the neural network to attend to both, low and high-frequency details, see [23] and [39] for further information.

Numerous advancements have been made to improve upon the original NeRF framework, notably focusing on accelerating both the training and inference times from days to near real-time [5, 14, 24, 45, 49]. One pivotal work in this regard is Instant-NGP [24], which revises the architecture by partitioning the single unified multi-layer perception (MLP) responsible for coarse and dense sampling into two distinct networks dedicated to density and color estimation. Additionally, Instant-NGP introduces *Multiresolution Hash Encoding* for input data, enabling the use of smaller MLPs and thereby speeding up the training process without compromising the rendering quality. In this paper, we adopt both, the architectural design and input encoding employed in Instant-NGP.

Multi-modal NeRFs. Integrating multi-modality into NeRFs is a fairly new field of research, and only a few works exist that try to combine different modalities. Most of the recent multi-modal NeRFs have been trained on RGB images and some kind of depth information, originating either from LiDAR scans [11, 28, 40, 50], RGB-D images [5, 13], or ToF data [22]. Moreover, there are two works that recently tried to build multi-modal NeRFs from RGB and near-infrared images. Based on computed camera poses, [11] first back-projects RGB and infrared images into 3D, yielding a coarse point cloud for both modalities, and then estimates relative transformations between sensors using point cloud registration. Using RGB camera poses computed from COLMAP [32], X-NeRF [27] *learns* relative poses to the infrared sensor during training, and leverages *Normalized Cross-Device Coordinates* to deal with different camera intrinsics. Both groups do not share their code and/or rely on custom datasets not publicly available, making a fair comparison difficult.

The majority of the methods mentioned above (including [11, 27]) implicitly assume a *shared* volume density across different modalities, effectively using the architecture proposed in the third strategy (RGB-X) that we explore in this paper. Furthermore, while previous works focus on how to simultaneously align *and* fuse multi-modal images into a unified neural scene representation, the focus of this work is on the latter. Essentially, we are interested in the following question: Assuming a multi-modal dataset with perfectly aligned images, what is the best way to integrate two imaging modalities into a neural scene representation?

3 Method

We present four different strategies of how to include an additional modality, other than RGB, into neural scene representations: (1) training a base model from scratch, (2) fine-tuning, (3) adding a second branch to the base model, and (4) adding an extra component for the second modality. Notably, while the last two strategies result in a *single*, multi-modal scene representation, the first two strategies always yield two separate models: one for RGB, and one for the second modality. A schematic overview of the four strategies is given in Fig. 1.

Throughout this work, we use Instant-NGP [24] as our base model. In brief, Instant-NGP's architecture comprises a density network composed of three fully-connected layers with 64 hidden dimensions each, as well as a color network featuring three fully-connected layers but with 32 hidden dimensions each. The density network takes hash-encoded coordinates as input and outputs a 16-dimensional vector, providing point-wise densities along with a 15-dimensional geometric descriptor. The color network processes the 15-dimensional descriptor in conjunction with an encoded viewing direction to generate view-dependent RGB color values. It is noteworthy that the density network's configuration remains consistent across all four strategies explained in the following.

Training from scratch (TS). In the first and simplest strategy, we train our base model from scratch, separately for RGB and the second modality. Since it is extremely challenging to compute reliable camera poses for thermal images using standard structure-from-motion approaches (see supp. material), we leverage poses derived from the corresponding RGB images and employ the following NeRF-like loss for training on thermal images:

$$\mathcal{L}_t = \sum_{\mathbf{r}\in\mathcal{R}} \left(\hat{t}(\mathbf{r}) - t(\mathbf{r})\right)^2, \tag{2}$$

where $\mathcal{R}$ is a set of rays, and $\hat{t}(\mathbf{r})$ and $t(\mathbf{r})$ are the predicted and ground-truth temperature values, respectively. Note that this loss is task-specific and may vary depending on the modality at hand. For training on RGB images, we use the standard rendering loss between predicted and true pixel color, $\hat{\mathbf{c}}(\mathbf{r})$ and $\mathbf{c}(\mathbf{r})$, respectively:

$$\mathcal{L}_c = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{\mathbf{c}}(\mathbf{r}) - \mathbf{c}(\mathbf{r})\|_2^2. \tag{3}$$

This strategy serves as our baseline.

Fine tuning (FT). Our second strategy first trains the base model on RGB images and then fine-tunes on images from the second modality, for which we again leverage RGB-derived camera poses. The idea behind this strategy is based on the assumption that the underlying scene's geometry is similar in both modalities and that training of the second modality can profit from being initialized with RGB data. For training on RGB images, we apply the same loss as in (3); similarly, fine-tuning on thermal images is done using the loss in (2).

Second branch (RGB-X). Our third strategy leverages multi-task learning by adding a second branch to the color network that predicts values of the second

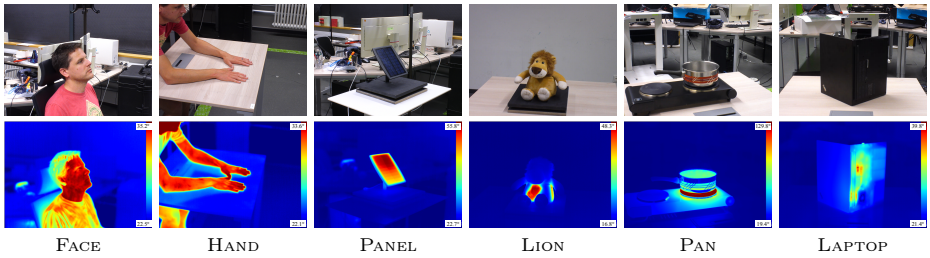

Fig. 2: Overview of our newly-captured dataset containing high-quality aligned RGB and thermal images of six common objects. Face, Hand, and Panel are forward-facing scenes consisting of around 40 images each. Lion, Pan, and Laptop are 360-degree scenes, where each scene has around 80 images.

modality. Although sharing a similar motivation as FT, RGB-X is the first of two proposed strategies that incorporate the additional modality into a single model. During training, we back-propagate both, RGB *and* predicted values of the second modality through the density network. Consequently, the density network is not only influenced by RGB but also the second modality. In our case, the color network is adapted to produce a four-dimensional RGBt output, comprising RGB color values *and* temperatures, t.

Separate component (SC). Contrary to utilizing a shared density for both modalities as in RGB-X, there are scenarios where this approach is sub-optimal for integrating diverse modalities. For instance, in our case, when using thermal and RGB imagery, the object's geometry under infrared deviates from its visible representation, please see supp. material. To rectify these disparities, and hence being able to predict accurate temperature values, especially for visible regions, it becomes imperative to leverage *only* RGB-derived densities. To account for this, our last strategy adds an extra component to the base model that solely predicts values of the second modality, but, contrary to RGB-X, *restricts* back-propagation to the density network during training. This constraint encourages the additional network to implicitly approximate RGB-derived geometry and prevents densities from being influenced by the second modality. We use a weighted combination of the previous loss functions from (2) and (3) for training on thermal images:

$$\mathcal{L} = \omega_c \mathcal{L}_c + \omega_t \mathcal{L}_t, \tag{4}$$

where we keep $w_c = w_t = 1$ constant (see supp. material for an ablation). The separate network predicting temperatures shares the same architecture as our base model's color network.

4 Dataset

We use a custom dataset containing RGB and thermal images of three forward-facing and three 360-degree scenes to compare previously explained strategies,

see Fig. 2. In total, our dataset, which we call *ThermalMix*, contains six common objects (Face, Hand, Panel, Lion, Pan, and Laptop) and is publicly available.

The data acquisition setup comprises a thermal camera (VarioCam HD, InfraTec GmbH, Germany) equipped with a 640×480 pixel resolution at 60 Hz for both, RGB and infrared sensor, a measurable temperature range of -40 to $+2,000$ degrees Celsius with an accuracy of ± 1 degree Celsius, and an infrared spectrum of 7.5 to 14 µm. Each of the six objects is placed on a table while the camera is moving around the object with a constant distance of about 1 m. Each forward-facing scene contains about 40 images, whereas about 80 images were taken for 360-degree scenes.

Cross-modality calibration. Since RGB and thermal images are captured from different sensors each one of them having its own coordinate system, a calibration object is positioned at the scene's center prior to data capturing, based on which we estimate the relative transformation between both sensors. Aligning images from different modalities is crucial in multi-modal reconstruction [11,27], and we are in need of a calibration target with easily identifiable features across both modalities. Finding the correct calibration object for RGB and thermal imagery, however, is a non-trivial task itself, see, *e.g.*, [34] or [38]. As the two modalities strongly differ in radiosity, common materials, objects, and patterns (*e.g.*, the classical checkerboard pattern printed on paper) can not be used, simply because they will not be visible in the infrared spectrum. Instead, we chose a perforated plate made out of aluminum as a calibration target, see supp. material. After slightly warming up (or cooling down) the plate, the holes are easily recognizable under both modalities, and we ultimately detect midpoints as matching features. Finally, since the distance between the camera and the object is fixed, we derive camera poses for thermal images from the previously computed transformation and absolute poses estimated from the corresponding RGB images (using COLMAP [32]). Please see supp. material for more details.

5 Results

Based on our *ThermalMix* dataset, we conducted extensive experiments to compare the proposed strategies. We start by quantitatively and qualitatively evaluating their ability to reconstruct (during training left out) thermal images in Section 5.2 and continue with presenting reconstruction results on RGB images in Section 5.3. Finally, we also investigate if our results generalize to other modalities, including near-infrared images and depth maps, in Section 5.4.

5.1 Implementation Details

Pre-processing. For pre-processing, RGB and thermal images are normalized within the range $[0, 1]$. Notably, for thermal images, normalization is performed relative to the *scene's* maximum temperature (*i.e.*, the maximum temperature

Table 1: Quantitative results on thermal images for (a) the three forward-facing scenes and (b) the three 360-degree scenes, measured using PSNR and SSIM (higher is better). Results were obtained from NeRFs trained on RGB and thermal data.

	TS		FT		RGB-t		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Face	30.34	**0.77**	30.04	0.75	**33.44**	0.68	32.10	0.66
Hand	35.54	**0.81**	33.99	0.73	**36.34**	0.73	33.56	0.60
Panel	31.21	**0.74**	29.66	0.55	**31.36**	0.61	27.31	0.38
	32.36	**0.77**	31.23	0.68	**33.71**	0.67	30.99	0.55

(a) Forward-facing scenes.

	TS		FT		RGB-t		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Lion	21.83	0.51	25.13	0.52	**27.82**	**0.61**	27.59	0.60
Pan	20.46	0.53	24.14	0.50	**27.48**	**0.54**	26.43	0.53
Laptop	23.15	0.37	24.95	0.49	**30.17**	**0.59**	28.07	0.53
	21.81	0.47	24.74	0.50	**28.49**	**0.58**	27.36	0.55

(b) 360-degree scenes.

over all images of a scene), as opposed to the conventional per-image maximum. This is necessary to ensure a consistent re-mapping after training.

Training. For each object, we use the same empirically determined parameters across all four strategies (clearly, while we use different parameter settings for each object, we keep parameters fixed across strategies). For FT, we pre-train for 6,000 iterations on RGB and fine-tune for another 4,000 iterations on thermal images. Networks of the remaining strategies (TS, RGB-X, and SC) are trained for 10,000 iterations each. Moreover, we used a batch of 4,096 rays per iteration, Adam [20] optimizer with default parameters, and a learning rate of 0.01.

Evaluation. Following the literature, we report Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) to evaluate our models, computed using leave-one-out cross-validation. Specifically, in each training process, a single image is set aside as a test sample while the model is trained on the remaining data. This procedure is repeated 10 times, and results were averaged. Finally, since our primary focus remains on the temperatures of the scene's central object, we segment objects within test images and compute evaluation metrics only in regions covered by an object (for training, however, we use the full, unsegmented images). Segmentation of thermal images is guided by RGB-derived segmentation masks.

5.2 Evaluation on Thermal Images

We present reconstruction results on thermal images separately for forward-facing and 360-degree scenes.

Forward-facing scenes. Results for forward-facing scenes (Face, Hand, and Panel) are shown in Table 1(a) and Fig. 3. As seen, RGB-X (which we denote as RGB-t in the following) outperforms all other strategies when considering PSNR. In terms of SSIM, however, TS performs best. This outcome is not surprising, considering the fact that TS solely relies on thermal measurements for its evaluations. More interestingly, we observe that RGB-t's performance in terms of SSIM is closest to the performance of FT (absolute deviation of 0.13; second-best is 0.29 between TS and FT). Since densities in RGB-t are affected by back-propagated temperature values, this suggests that the RGB-t density network somehow balances between RGB and thermal densities, allowing FT

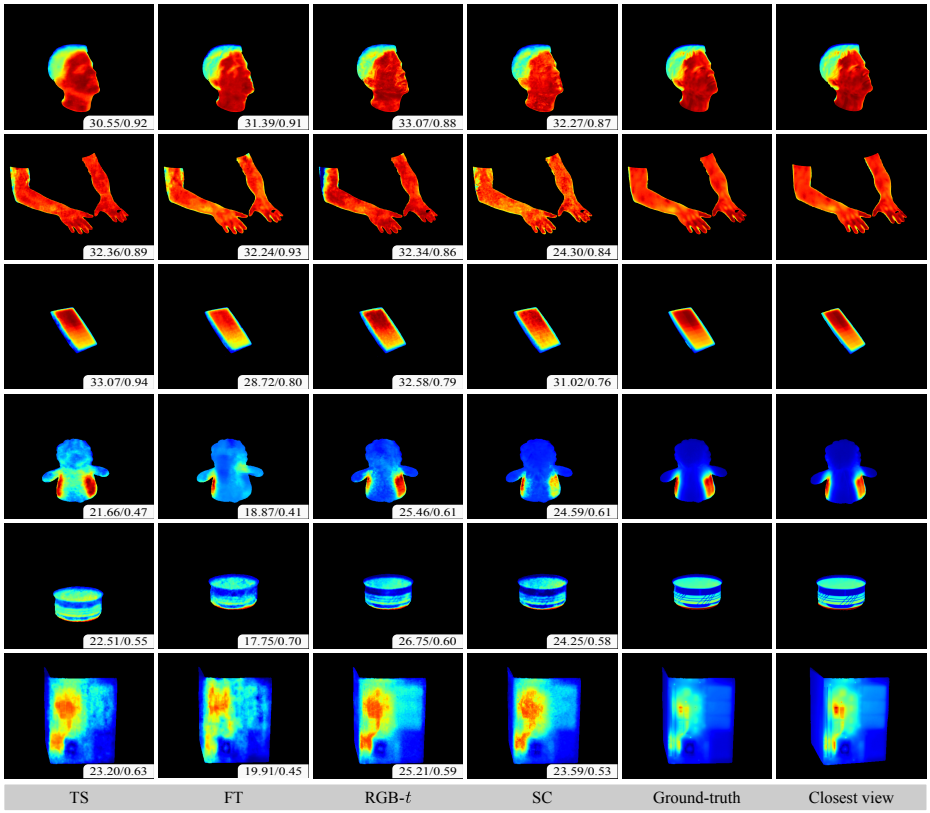

Fig. 3: Reconstructions of a (left-out) thermal image from multi-modal neural scene representations trained on RGB and thermal data, arising from the four strategies that we compare. For each view, we also report PSNR and SSIM (higher is better). *Closest view* denotes the nearest image in the training set.

to further lean on the thermal densities during fine-tuning. Furthermore, our empirical evaluation shows that the performance of SC seems to be inconsistent, varying with the object being reconstructed. This variability is understandable given that SC does not integrate thermal densities, making the complexity of the scene a critical factor in the performance of the separate thermal network.

360-degree scenes. We report PSNRs and SSIMs for 360-degree scenes (Lion, Pan, and Laptop) in Table 1(b) and qualitative results in Fig. 3. Contrary to forward-facing scenes, RGB-t outperforms other strategies in both, PSNR *and* SSIM for 360-degree scenes. Also, another key observation is the minimal variation in both metrics across different objects, unlike other strategies. This uniformity demonstrates the robustness of RGB-t against the challenges inherent in thermal imagery and underscores the significance of using RGB images to guide thermal reconstruction.

Table 2: Quantitative results on RGB images, measured using PSNR and SSIM (higher is better). Results were obtained from NeRFs trained on RGB and thermal data. FT is left out since its RGB component is similar to TS.

	TS		RGB-t		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Face	**30.78**	**0.85**	29.46	0.79	30.12	0.82
Hand	**30.81**	**0.93**	30.37	0.82	30.62	0.88
Panel	**30.56**	**0.84**	29.81	0.79	30.03	0.80
Lion	**30.71**	**0.82**	29.08	0.73	30.18	0.77
Pan	**29.59**	**0.76**	29.33	0.73	29.40	0.72
Laptop	**29.70**	**0.74**	29.42	0.72	29.45	0.72
	30.36	**0.82**	29.58	0.76	29.67	0.79

Moreover, we observe that TS and FT perform poorly with regard to PSNR, showing that these strategies exhibit serious difficulties in achieving high-quality results on thermal images within a 360-degree context. A possible explanation for this effect may be the static nature of the background in thermal images, which tends to not have as much variation as seen in RGB images. Ultimately, this introduces ambiguities, effectively hampering the network to distinguish between forward- and backward-facing views (this becomes even worse if the object itself is rotationally symmetrical, such as the pan, for instance). Finally, we would like to note that TS, unlike FT, seems to be more sensitive to floaters. Please refer to the supp. material for more information.

The performance of SC consistently ranks second to RGB-t (both, for PSNR and SSIM), reinforcing the notion that relying solely on thermal images in 360-degree scenes can yield subpar results, primarily due to the near-static background. In SC, where only the densities from the RGB data are utilized, we observe an improvement over TS but results are either comparable or inferior to RGB-t. This outcome highlights the limitations of using only RGB densities in a thermal context and underlines the added value of incorporating RGB information directly, as done in the RGB-t.

5.3 Evaluation on RGB Images

Ideally, a multi-modal neural scene representation should be able to reconstruct images at least as good as its uni-modal counterpart. In this section, we will assess the performance of the proposed strategies on RGB images. Since FT's RGB component is similar to TS (they only differ in the number of iterations trained), we only show results for TS, RGB-t, and SC in this section.

Quantitative and qualitative results can be found in Table 2 and Fig. 4. As can be observed, our baseline strategy, TS, *slightly* outperforms all the other strategies in both, PSNR and SSIM. Moreover, SC maintains superior reconstruction quality in terms of PSNR and SSIM compared to RGB-t. This result can be attributed to SC's non-interference with RGB densities, whereas RGB-t integrates thermal and RGB densities, causing a mixture of information. Ultimately, when comparing both strategies to TS (which was solely trained on

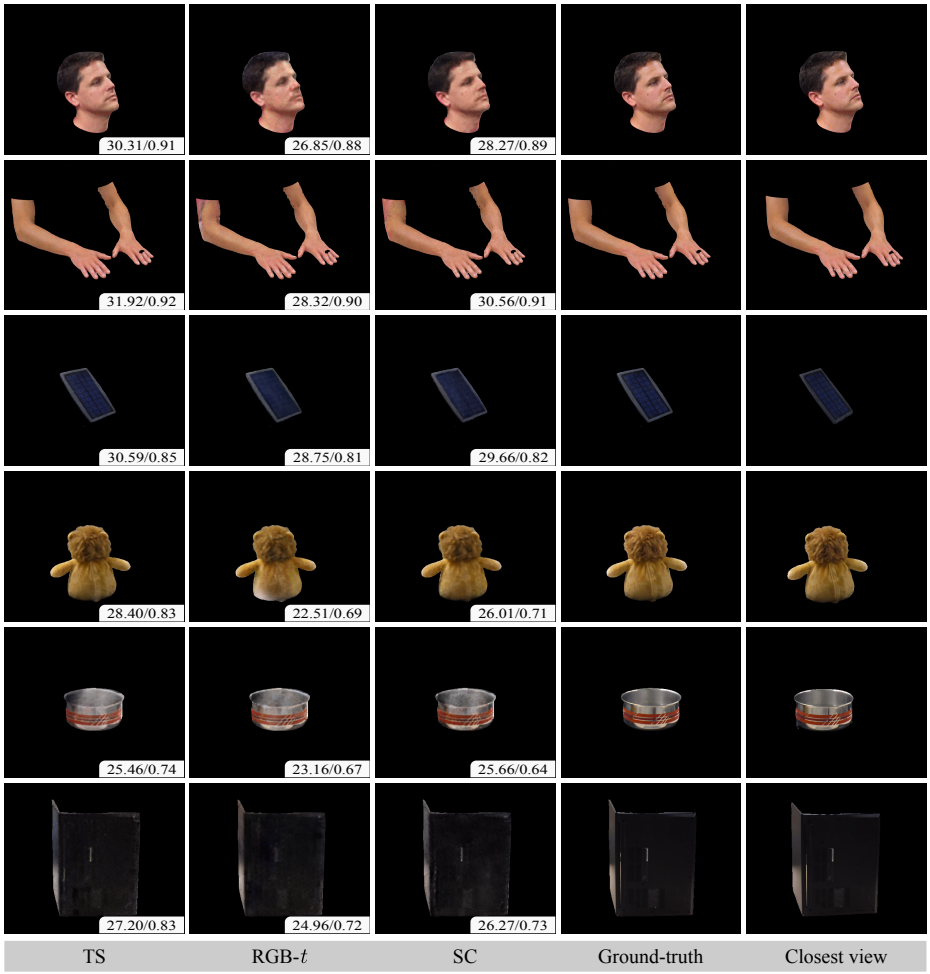

Fig. 4: Reconstructions of a (left-out) RGB image from multi-modal neural scene representations trained on RGB and thermal data, arising from the four strategies that we compare. FT is left out since its RGB component is similar to TS. For each view, we also report PSNR and SSIM (higher is better). *Closest view* denotes the nearest image in the training set.

Table 3: Quantitative results on (a) NIR and (b) RGB images, measured using PSNR and SSIM (higher is better). Results were obtained from NeRFs trained on RGB and NIR data. FT is left out since its RGB component is similar to TS.

	TS		FT		RGB-NIR		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Snail	36.55	**0.97**	32.32	0.95	**36.70**	**0.97**	35.55	0.96
Bear	**37.15**	**0.97**	35.01	0.95	37.01	0.96	36.05	0.96
Elephant	**35.11**	**0.97**	33.60	0.96	35.09	**0.97**	34.99	0.95
	36.27	**0.97**	33.64	0.95	**36.27**	**0.97**	35.53	0.96

(a) NIR reconstruction quality.

	TS		RGB-NIR		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Snail	**39.94**	**0.97**	37.31	0.96	38.03	**0.97**
Bear	**38.10**	**0.96**	36.29	**0.96**	37.93	**0.96**
Elephant	**39.71**	**0.97**	38.07	0.96	39.10	**0.97**
	39.25	**0.97**	37.22	0.96	38.35	**0.97**

(b) RGB reconstruction quality.

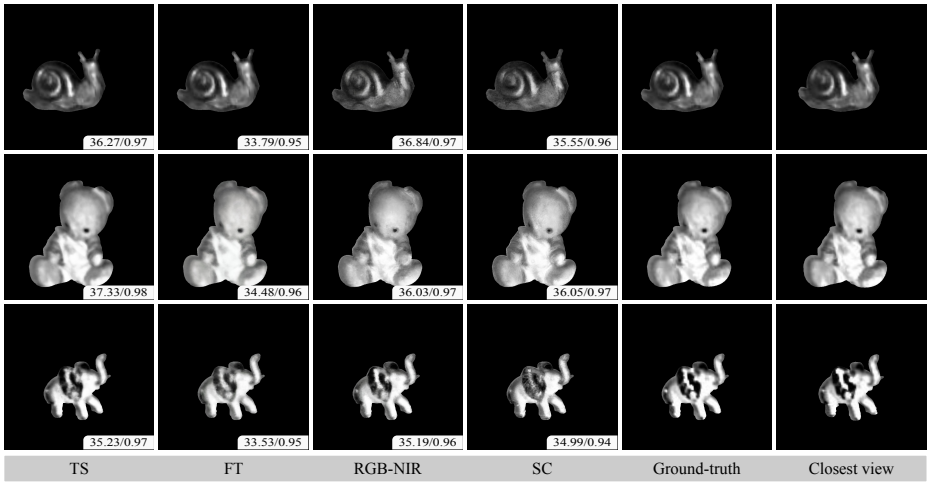

Fig. 5: Reconstructions of a (left-out) NIR image from multi-modal neural scene representations trained on RGB and NIR data, arising from the four strategies that we compare. For each view, we also report PSNR and SSIM (higher is better). *Closest view* denotes the nearest image in the training set.

RGB images), we find that SC achieves similar reconstruction quality, whereas RGB-t lags slightly behind. Note that those findings also match the presented qualitative results, where only small differences between TS, RGB-t, and SC can be observed with the human eye.

5.4 Evaluation on Other Modalities

Lastly, we also report results on multi-modal NeRFs learnt from RGB and near-infrared (NIR) images, and RGB and depth maps. Due to space constraints, results for the latter can be found in the supp. material. However, they match exactly what we present here.

For all experiments, we used three randomly selected forward-facing scenes from the multi-sensor dataset proposed in [43]. Each scene contains 100 images, and we took images captured with the Huawei Mate 30 Pro (with a resolution

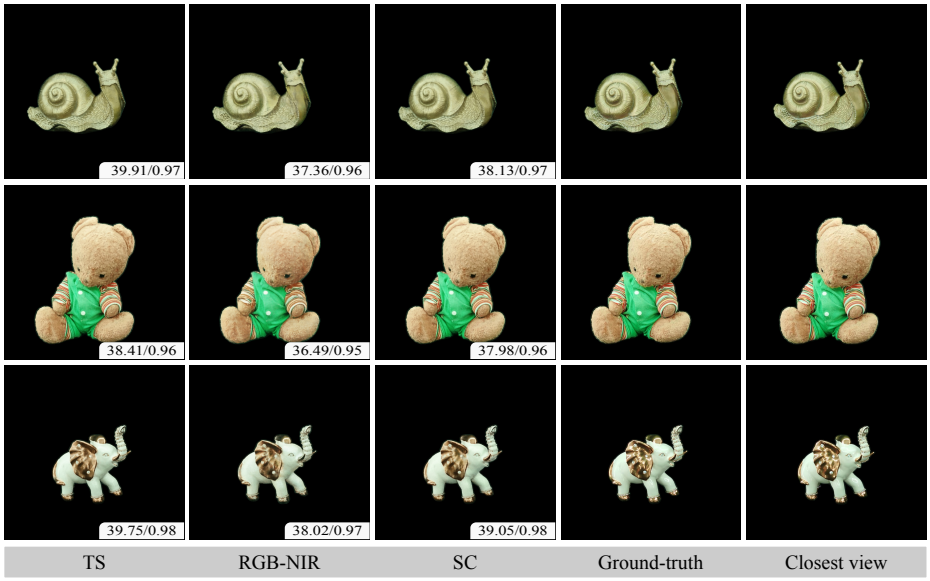

Fig. 6: Reconstructions of a (left-out) RGB image from multi-modal neural scene representations trained on RGB and NIR data, arising from the four strategies that we compare. FT is left out since its RGB component is similar to TS. For each view, we also report PSNR and SSIM (higher is better). *Closest view* denotes the nearest image in the training set.

of 7296×5472 for RGB and 240×180 for NIR images). Note that, due to the different resolutions, a direct comparison between results obtained from RGB and thermal (see Sections 5.2 and 5.3) and RGB and NIR would be unfair. However, since we are anyway only interested in a relative comparison between proposed strategies, having different resolutions is not a problem.

Evaluation on NIR images. We report results in Table 3(a) and Fig. 5. Considering NIR reconstruction quality, RGB-X (denoted as RGB-NIR in the following) performs best on average, but, as opposed to RGB and thermal, this time on par with TS. This is understandable given the fact that NIR images do not have an as static and texture-less background as thermal images, effectively mitigating the problems described in Section 5.2. All in all, we observe that incorporating NIR images into a multi-modal NeRF seem to be easier (*i.e.*, less dependent on the employed strategy), which (i) confirms our hypothesis that modeling thermal images seem to be hard, and (ii) justifies the usage of thermal imagery as a challenging benchmark for evaluating the proposed strategies.

Evaluation on RGB images. Quantitative and qualitative results on RGB reconstruction quality can be found in Table 3(b) and Fig. 6. We observe a very similar trend as for multi-modal NeRFs learnt from RGB and thermal images: TS performs best in terms of both, PSNR and SSIM, followed by SC and RGB-NIR.

6 Discussion and Limitations

Based on our analysis, RGB-X consistently outperforms all strategies in reconstructing thermal images, and performs on par with our baseline, TS, when it comes to NIR images and depth maps. Taking into account RGB reconstruction quality, SC slightly surpasses RGB-X when trained on RGB and thermal, NIR, or depth data. Moreover, our analysis reveals that, especially for low-texture images such as thermal images, allowing NeRF's volume densities to be influenced by the second modality (as in RGB-X) helps a lot in learning multi-modal neural scene representations, effectively causing a mixture of information. All in all, due to the fact that RGB-X still yields compelling results on RGB data (although it can not meet the quality of its uni-modal counterpart), we conclude that RGB-X seems to be well suited for multi-modal neural scene representations.

Limitations. We also want to note a limitation of the present study, predominantly from a practical point of view. First off, although offline cross-modality calibration provides almost perfect alignments between RGB and thermal images, it prevents us from building NeRFs of uncalibrated, in-the-wild images. This could be approached by integrating learning-based calibration methods as proposed in [11, 27]. However, since the focus of this work is clearly *not* on (learning) cross-modality calibration but rather on the systematic evaluation of different strategies to integrate multi-modality into NeRFs, we leave this for future work.

7 Conclusion and Future Work

In this paper, we have systematically compared four different strategies of how to include image data from different modalities, other than RGB, into a single scene representation, ultimately aiming for multi-modal neural scene representations. Based on RGB data and a NeRF-like scene representation as our base model, we propose to include images from a second modality using (1) training from scratch (TS), (2) fine-tuning (FT), (3) adding a second branch (RGB-X), and (4) adding a separate component (SC) to the base model. The analysis of the four strategies is based on a newly captured, object-centric dataset, named *ThermalMix*, which consists of 360 multi-view RGB and thermal images. The dataset includes six common objects in total, three of them captured as forward-facing scenes and three as 360-degree scenes, and it is the first to provide near-perfect alignments between RGB and thermal images. *ThermalMix* is publicly available, and we believe it serves as a challenging benchmark not only for reconstruction tasks, but also for learnable cross-modality calibration.

Our findings indicate that RGB-X stands out for its thermal reconstruction capabilities while also delivering compelling RGB reconstructions. Finally, we could also show that our results generalize to other modalities, including NIR images and depth maps, leading to the conclusion that RGB-X indeed seems to be well-suited for building general multi-modal neural scene representations.

Future work. Our future work is primarily concerned with the incorporation and evaluation of learning-based schemes for online cross-modality calibration.

Acknowledgments. We would like to thank Ian Marius Peters, Bernd Doll, and Oleksandr Mashkov for valuable discussions and access to the thermal camera. This work was funded by the German Federal Ministry of Education and Research (BMBF), FKZ: 01IS22082 (IRRW). The authors are responsible for the content of this publication.

References

1. Aggarwal, A.K.: Thermal imaging for cancer detection. Imaging Radiat Res **6**, 1–13 (2023) 3
2. Akula, A., Ghosh, R., Sardana, H.K.: Thermal imaging and its application in defence systems. In: AIP Conf Proc (2011) 3
3. Alldieck, T., Bahnsen, C.H., Moeslund, T.B.: Context-aware fusion of rgb and thermal imagery for traffic monitoring. Sensors **16** (2016) 3
4. Ceyhun, K., Ozgun, P., Sultan, T.: 3d mesh model generation from ct and mri data. In: IEEE BigData 2021. pp. 4725–4730 (2021) 2
5. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: CVPR. pp. 12882–12891 (2021) 2, 4
6. Dongdong, M., Sheng, L., Bin, S., Hao, W., Suqing, T., Wenjun, M., Guoping, W., Xueqing, Y.: 3d reconstruction-oriented fully automatic multi-modal tumor segmentation by dual attention-guided vnet. Vis Comput **39**, 3183–3196 (2023) 2
7. Gade, R., Moeslund, T.B.: Thermal cameras and applications: A survey. Mach Vis Appl (2014) 3
8. Gaetano, M., Giuseppe, M.: Applications of uav thermal imagery in precision agriculture: State of the art and future research outlook. Remote Sens **12** (2020) 3
9. Gao, K., Gao, Y., He, H., Lu, D., Xu, L., Li, J.: Nerf: Neural radiance field in 3d vision, a comprehensive review. arXiv:2210.00379 (2023) 2
10. Gowen, A.A., Tiwari, B.K., Cullen, P.J., McDonnell, K., O'Donnell, C.P.: Applications of thermal imaging in food quality and safety assessment. Trends Food Sci **21**, 190–200 (2010) 3
11. Haidong, Z., Yuyin, S., Chi, L., Lu, X., Jiajia, L., Nan, Q., Ramkant, N., Cheng-Hao, K.: Multimodal neural radiance field. In: ICRA. pp. 9393–9399 (2023) 2, 3, 4, 7, 14
12. Han, W., Liu, X., Song, S., Meng, M.Q.H.: 3d reconstruction of dense model based on the sparse frames using rgbd camera. In: ROBIO. pp. 2726–2731 (2019) 2
13. Haoyi, Z.: X-nerf: Explicit neural radiance field for multi-scene 360° insufficient rgb-d views. In: WACV. pp. 5766–5775 (2023) 2, 4
14. Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P.: Baking neural radiance fields for real-time view synthesis. ICCV pp. 5875–5884 (2021) 4
15. Huo, D., Wang, J., Qian, Y., Yang, Y.H.: Glass segmentation with rgb-thermal image pairs. IEEE Trans Image Process **32**, 1911–1926 (2023) 3
16. Ishimwe, R., Abutaleb, K., Ahmed, F.: Applications of thermal imaging in agriculture–a review. ARS **3**, 128–140 (2014) 3
17. Jones, H.G.: Application of thermal imaging and infrared sensing in plant physiology and ecophysiology. Adv Bot Res **41**, 107–163 (2004) 3
18. Joshi, N.P., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Rudbeck, M.J., Kuemmerle, T., Meyfroidt, P., Mitchard, E.T.A., Reiche, J., Ryan, C.M., Waske, B.: A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. Remote Sens **8** (2016) 2

19. Khanal, S., Fulton, J.P., Shearer, S.A.: An overview of current and potential applications of thermal remote sensing in precision agriculture. Comput Electron Agric **139**, 22–32 (2017) 3
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: CoRR (2014) 8
21. Li, G., Lin, Y., Ouyang, D., Li, S., Luo, X., Qu, X., Pi, D., Li, S.E.: A rgb-thermal image segmentation method based on parameter sharing and attention fusion for safe autonomous driving. IEEE Trans Intell Transp Syst pp. 1–16 (2023) 3
22. Liu, X., Li, Y., Teng, Y., Bao, H., Zhang, G., Zhang, Y., Cui, Z.: Multi-modal neural radiance field for monocular dense slam with a light-weight tof sensor. In: ICCV (2023) 2, 4
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 1, 3, 4
24. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans Graph **41** (2022) 2, 4, 5
25. Örs Petneházy, Rück, S., Sós, E., Reinitz, L.Z.: 3d reconstruction of the blood supply in an elephant's forefoot using fused ct and mri sequences. Animals **13** (2023) 2
26. Pineda, M., Barón, M., Pérez-Bueno, M.L.: Thermal imaging for plant stress detection and phenotyping. Remote Sens **13** (2021) 3
27. Poggi, M., Ramirez, P.Z., Tosi, F., Salti, S., Mattoccia, S., Stefano, L.D.: Cross-spectral neural radiance fields. In: I3DV (2022) 2, 3, 4, 7, 14
28. Quentin, H., Nathan, P., Moussâb, B., Luis, R., D., T., C., M., P., V., C., D.: Moisst: Multi-modal optimization of implicit scene for spatiotemporal calibration. In: IROS (2023) 2, 4
29. Rai, M.K., Maity, T., Yadav, R.K.: Thermal imaging system and its real time applications: a survey. J Eng Technol **25**, 245–262 (2017) 3
30. Raju, V.B., Imtiaz, M.H., Sazonov, E.: Food image segmentation using multi-modal imaging sensors with color and thermal data. Sensors **23** (2023) 3
31. Ring, E.F.J., Ammer, K.: Infrared thermal imaging in medicine. Physiol Meas **33**, 33–46 (2012) 3
32. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016) 1, 4, 7, 18, 19
33. Shaikh, S., Akhter, N., Manza, R.: Current trends in the application of thermal imaging in medical condition analysis. IJITEE **8**, 2708–2712 (2019) 3
34. Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J.: Pst900: Rgb-thermal calibration, dataset and segmentation network. In: ICRA. pp. 9441–9447 (2020) 3, 7
35. Shum, H.Y., Kang, S.B.: Review of image-based rendering techniques. In: VCIP. vol. 4067, pp. 2–13 (2000) 1
36. de Souza, M.A., Cordeiro, D.C.A., de Oliveira, J., de Oliveira, M.F.A., Bonafini, B.L.: 3d multi-modality medical imaging: Combining anatomical and infrared thermal images for 3d reconstruction. Sensors **23** (2023) 3
37. Still, C.J., Powell, R.L., Aubrecht, D.M., Kim, Y., Helliker, B.R., Roberts, D.A., Richardson, A.D., Goulden, M.L.: Thermal imaging in plant and ecosystem ecology: applications and challenges. Ecosphere **10** (2019) 3
38. Swamidoss, I.N., Amro, A.B., Sayadi, S.: Systematic approach for thermal imaging camera calibration for machine vision applications. Optik **247** (2021) 7

39. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: NIPS (2020) 4
40. Tang, T., Wang, G., Lao, Y., Chen, P., Liu, J., Lin, L., Yu, K., Liang, X.: Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis. arXiv:2402.17483 (2024) 2, 4
41. Tayebi, R.M., Wirza, R., Sulaiman, P.S., Dimon, M.Z., Khalid, F., Al-Surmi, A.A., Mazaheri, S.: 3d multimodal cardiac data reconstruction using angiography and computerized tomographic angiography registration. J Cardiothorac Surg **10** (2015) 2
42. Vadivambal, R., Jayas, D.S.: Applications of thermal imaging in agriculture and food industry – a review. Food Bioproc Tech **4**, 186–199 (2011) 3
43. Voynov, O., Bobrovskikh, G., Karpyshev, P., Galochkin, S., Ardelean, A.T., Bozhenko, A., Karmanova, E., Kopanev, P., Labutin-Rymsho, Y., Rakhimov, R., Safin, A., Serpiva, V., Artemov, A., Burnaev, E., Tsetserukou, D., Zorin, D.: Multi-sensor large-scale dataset for multi-view 3d reconstruction. In: CVPR. pp. 21392–21403 (2023) 12
44. Wakeford, Z.E., Chmielewska, M., Hole, M.J., Howell, J.A., Jerram, D.A.: Combining thermal imaging with photogrammetry of an active volcano using uav: an example from stromboli, italy. The Photogrammetric Record **34** (2019) 2
45. Wang, D., Zhang, T., Abboud, A., Süsstrunk, S.: Inpaintnerf360: Text-guided 3d inpainting on unbounded neural radiance fields. arXiv:2305.15094 (2023) 4
46. Wilson, A.N., Gupta, K., Koduru, B.H., Kumar, A., Jha, A., Cenkeramaddi, L.R.: Recent advances in thermal imaging and its applications using machine learning: A review. IEEE Sens J **23**, 3395–3407 (2023) 3
47. Yang, X., Guo, R., Li, H.: Comparison of multimodal rgb-thermal fusion techniques for exterior wall multi-defect detection. JIIR **2** (2023) 3
48. Yasutaka, F., Carlos, H.: Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision **9**, 1–148 (2015) 1
49. Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR. pp. 5501–5510 (2022) 4
50. Zhang, Q., Wang, B.H., Yang, M.C., Zou, H.: Mmnerf: Multi-modal and multi-view optimized cross-scene neural radiance fields. IEEE Access **11**, 27401–27413 (2023) 2, 4
51. Zollhöfer, M., Stotko, P., Gorlitz, A., Theobalt, C., Nießner, M., Klein, R., Kolb, A.: State of the art on 3d reconstruction with rgb-d cameras. Computer Graphics Forum **37**, 625–652 (2018) 2
52. Štumper, M., Kraus, J.: Thermal imaging in aviation. MAD **3** (2015) 3

Exploring Multi-modal Neural Scene Representations With Applications on Thermal Imaging

— Supplementary Material —

Mert Özer, Maximilian Weiherer, Martin Hundhausen, and Bernhard Egger

Friedrich-Alexander-Universität Erlangen-Nürnberg
firstname.lastname@fau.de

In this supplementary material, we (i) demonstrate that it is challenging to compute (reliable) camera poses from thermal images (Section A), (ii) investigate the difference in geometry between RGB and thermal images (Section B), (iii) provide an ablation on the weights w_c and w_t in Eq. (4) of the main paper (Section C), (iv) give more details on how we align RGB and thermal images during cross-modality calibration (Section D), (v) further analyze why 360-degree scenes are harder to optimize than forward-facing scenes (Section E), and (vi) present results on multi-modal NeRFs learned from RGB and depth maps (Section F).

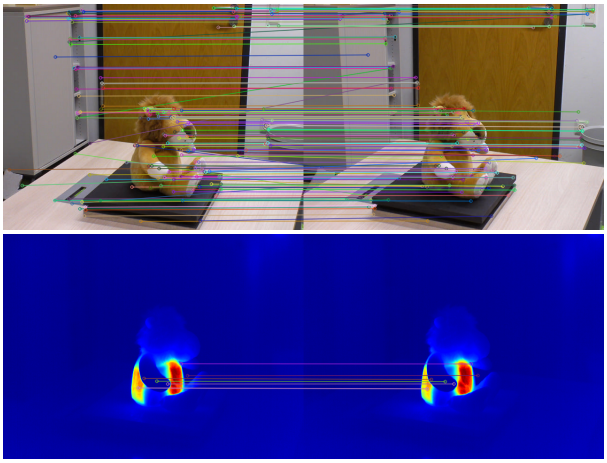

Fig. 1: Demonstration of how challenging it is to compute *reliable* camera poses from thermal images. We visualize feature correspondences between two views on RGB (top row) and thermal images (bottom row; found and matched using COLMAP [32]).

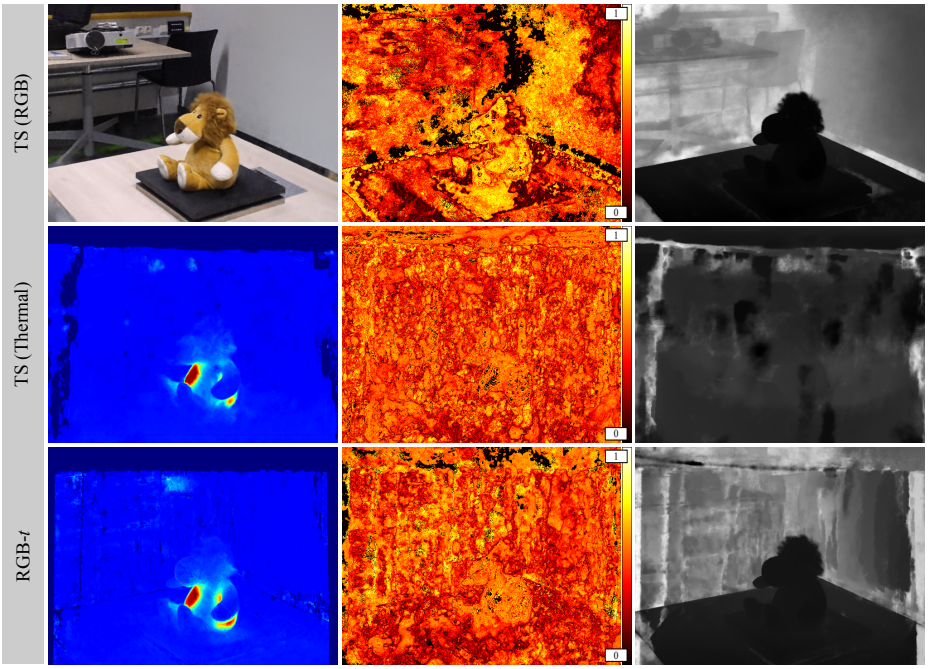

Fig. 2: Comparison of the reconstructed geometry in RGB (first row) and thermal images (second and third row), shown on TS and RGB-t. The first column shows novel renderings, the second column visualizes accumulated densities for each pixel along its respective ray, and the third column depicts estimated depth maps. As can be observed clearly, thermal-derived geometry greatly benefits from utilizing RGB densities (especially seen in the depth maps arising from TS trained solely on thermal images (second row) and depth maps produced by RGB-t (third row), which, contrary to TS, incorporates RGB information).

A Camera Poses From Thermal Images

In Fig. 1 we show matching features between two RGB images and corresponding thermal images found using COLMAP [32]. For RGB, 127 reliable matches have been found, while only seven matches could be found on thermal images, rendering it almost impossible to compute a meaningful camera pose. This demonstrates that classical feature extractors (and hence, structure-from-motion techniques) struggle with thermal images, ultimately requiring new, specialized methods for camera pose estimation from non-RGB images.

B Different Geometry From RGB and Thermal Images

As noted in Section 3 of the main paper and based on our analysis, we observe significant differences in scene densities when comparing RGB and thermal

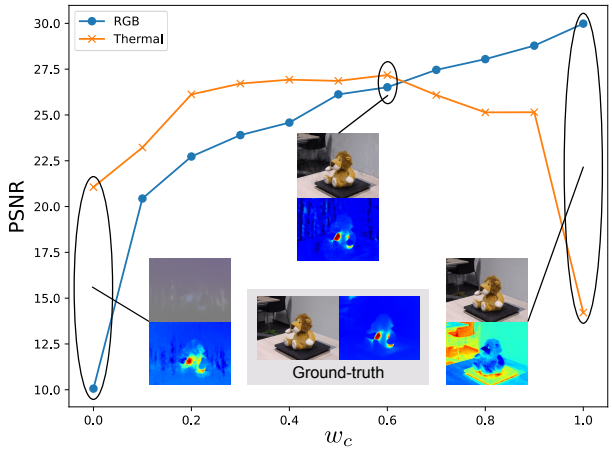

Fig. 3: Ablation on the weights w_c and w_t of Eq. (4) in the main paper. We evaluate for $w_c \in \{0, 0.1, 0.2, \ldots, 1\}$ and set $w_t = 1 - w_c$. Models were trained with RGB-t.

modalities, leading to variations in the reconstructed geometry of the scene. As illustrated in Fig. 2, the second column presents the accumulated densities for each pixel along its respective ray. This visualization highlights a distinct shift when transitioning from training exclusively with RGB data to training with thermal data from scratch (as done in TS). Notably, RGB-t appears to strike a balance in these density distributions. It is evident that thermal images can benefit from utilizing RGB densities, leading to improved reconstruction quality.

Further insights are offered in the third column showing predicted depth maps. As seen, depth maps elucidate that in thermal imaging, the uniform background tends to create the illusion of a single continuous surface across most of the image, except for the high-temperature regions on the object of interest. This effect underscores the unique challenges and considerations when interpreting thermal imagery in contrast to RGB and the advantage of integrating RGB information for more effective reconstruction of thermal data.

C Ablation on Weights of Loss

We provide an ablation on w_c and w_t as used in Eq. (4) of the main paper in Fig. 3. Results were obtained using RGB-t trained on the 360-degree Lion scene. As can be observed, values of the two weights matter, but are not very sensitive.

D Details on Cross-Modality Calibration

As explained in Section 4 of the main paper, we align RGB and thermal images using a perforated aluminum plate (which we just bought in a local hardware store) as calibration object visible in both imaging modalities, see Fig. 4.

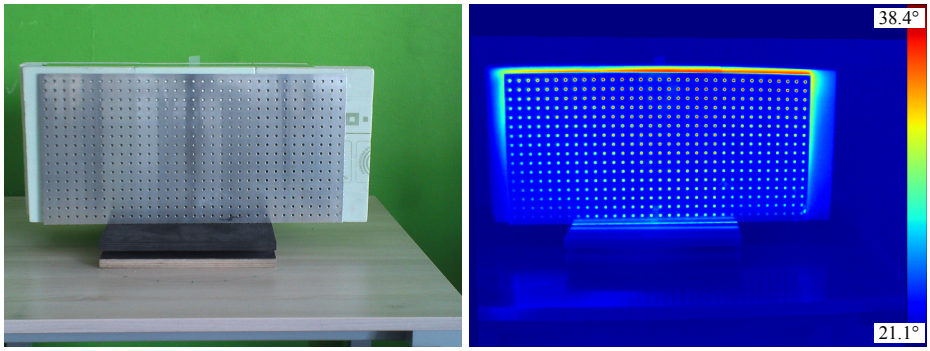

Fig. 4: Calibration object used to align RGB and thermal images. We utilized a perforated, slightly warmed-up aluminum plate, which is visible in both, RGB and thermal images. To compute relative poses between RGB and infrared sensors, we identify the holes' midpoints as matching features across the two modalities.

We establish at least four point correspondences (as stated, we detect midpoints of the holes) between RGB and thermal images of the calibration object. For corresponding points (x_i, y_i) in the RGB image and (x'_i, y'_i) in the thermal image, we calculate the homography matrix H using the following steps.

1. **Formulating equations:** Each set of corresponding points yields two equations:

$$x'_i = \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}}, \quad y'_i = \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}}.$$

2. **Rearranging into linear system:** These equations are rearranged to a linear form $Ax = b$, leading to the matrix equation:

$$\begin{bmatrix} x_i & y_i & 1 & 0 & 0 & 0 & -x'_i x_i & -x'_i y_i \\ 0 & 0 & 0 & x_i & y_i & 1 & -y'_i x_i & -y'_i y_i \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{bmatrix} = \begin{bmatrix} x'_i \\ y'_i \end{bmatrix}.$$

3. **Solving with SVD:** The system is typically over-determined, so we apply Singular Value Decomposition (SVD) to find the solution. The homography matrix H is obtained from the last column of V (from the decomposition $A = U\Sigma V^T$) corresponding to the smallest singular value. The resulting matrix H enables the transformation of coordinates from the RGB image plane to the thermal image plane, facilitating the accurate mapping of thermal images onto the RGB camera space.

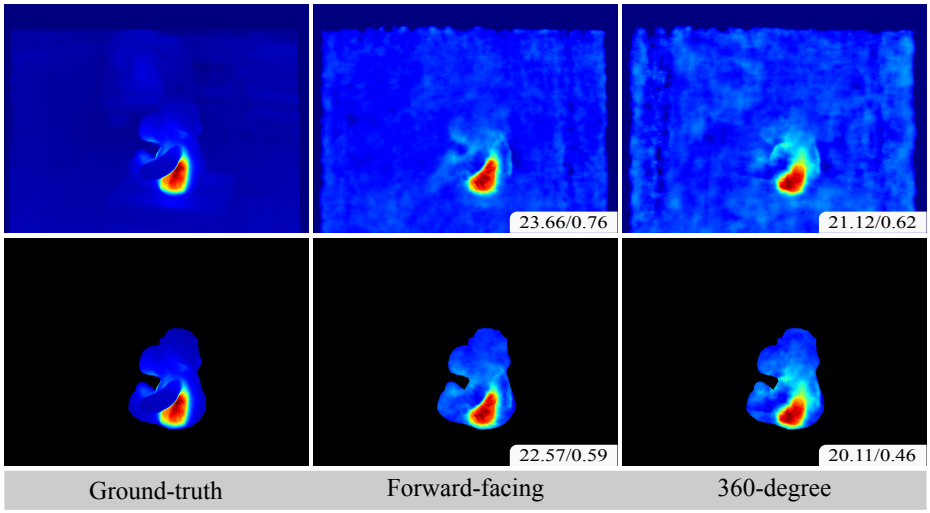

Fig. 5: Novel views (second and third column) produced by TS trained on LION. For the second column, we trained TS only on the forward-facing views from the dataset, while for the third column, we used all available images. Second row shows segmented predictions from the first row. We also report PSNR and SSIM (higher is better).

E Challenges on 360-Degree Scenes

Training NeRF on 360-degree scenes using thermal images presents unique complexities as noted in Section 5.2 of the main paper. One primary challenge is the static appearance of thermal backgrounds, particularly in object-centered scenarios. In such cases, the background often appears uniform, creating an illusion that the object is rotating rather than the camera moving around the object. This phenomenon can mislead the network during training, allowing it to minimize loss by simply adjusting the background color. In an experiment, where TS was trained on half and the entirety of a 360-degree LION scene for 10,000 iterations (as shown in Fig. 5), we observed that forward-facing training significantly reduces the occurrence of floaters present in the scene. One plausible reason why forward-facing scenes may be more effective than 360-degree scenes for depth estimation is the nature of background representation. In 360-degree scenes, the background often appears as a static surface, very similar at both the front and back of the central object, akin to a curtain enveloping the object (see Fig. 5). This uniformity can obscure the object, making depth perception challenging. In contrast, forward-facing scenes typically exhibit this static background effect only behind the object, not in front, allowing for clearer distinction and depth estimation of the central subject. Once again, this idea leads to the importance of RGB information utilization for thermal reconstruction.

Another issue arises with objects that have symmetrical shapes, such as PAN in our work. Both, the front and back views of such objects display very similar backgrounds in thermal imaging due to its nature. This similarity in views,

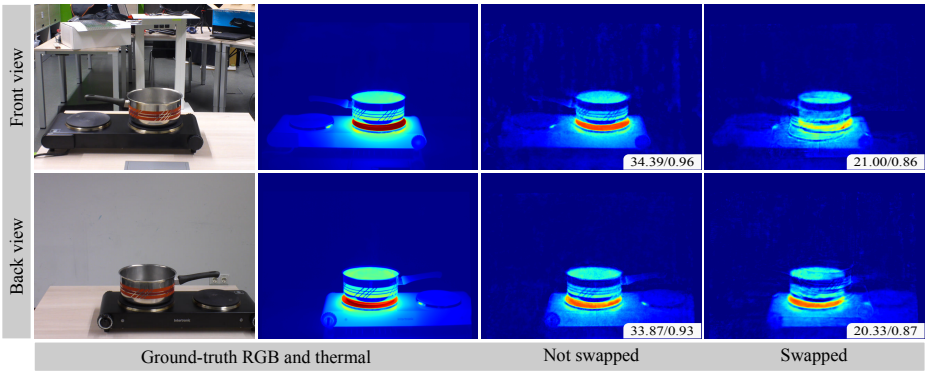

Fig. 6: Training on 360-degree scenes with symmetrical objects, such as PAN, is challenging in the context of thermal imaging. In this experiment, we trained RGB-*t* on two datasets: the original dataset, and a dataset in which we exchanged front and back views. We show a novel front-facing view in the first row and a back-facing view in the second row. We also report PSNR and SSIM (higher is better).

despite being 180 degrees apart in viewing direction, can cause ambiguities in training. To investigate this, we swapped the front and back views of the pan object and trained RGB-*t* for it. We found that the results remained satisfactory, as demonstrated in Fig. 6.

Table 1: Quantitative results on (a) depth maps and (b) RGB images, measured using PSNR and SSIM (higher is better). Results were obtained from NeRFs trained on RGB images and depth maps. FT is left out since its RGB component is similar to TS.

	TS		FT		RGB-D		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SNAIL	33.38	**0.94**	30.56	0.92	**33.51**	**0.94**	33.07	**0.94**
BEAR	**34.56**	**0.95**	31.78	0.94	34.09	0.94	33.82	0.94
ELEPHANT	**32.11**	**0.94**	30.81	0.92	32.08	0.93	31.95	**0.94**
	33.35	**0.94**	31.05	0.93	33.23	**0.94**	32.95	**0.94**

(a) Depth reconstruction quality.

	TS		RGB-D		SC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SNAIL	**39.94**	**0.97**	36.05	0.96	37.66	**0.97**
BEAR	**38.10**	**0.96**	35.60	0.94	37.78	**0.96**
ELEPHANT	**39.71**	**0.97**	36.86	0.95	38.39	**0.97**
	39.25	**0.97**	36.17	0.95	37.94	**0.97**

(b) RGB reconstruction quality.

F Results on RGB and Depth Maps

Finally, we present results obtained from multi-modal NeRFs trained on RGB and depth maps for all of the four strategies. We follow the same protocol as described in the main paper, see Section 5.4.

Quantitative and qualitative results can be found in Table 1 and Figs. 7 and 8. Note that those findings match what we have seen in other modalities. Specifically, considering depth reconstruction quality, we observe that TS and RGB-X (RGB-D in this case) perform almost on par, exhibiting similar behavior as for NeRFs trained on RGB and NIR images. The same is true for RGB reconstruction quality; here, TS performs best, followed by SC and RGB-D.

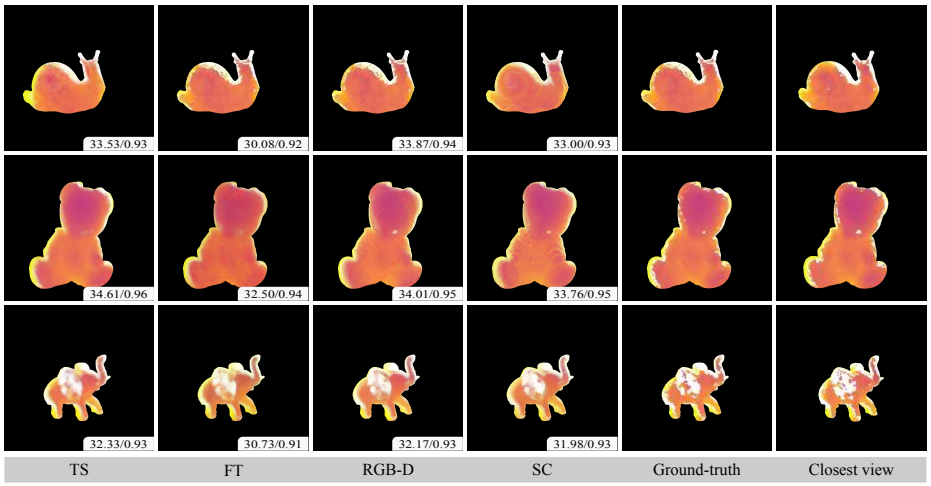

Fig. 7: Reconstructions of a (left-out) depth map from multi-modal neural scene representations trained on RGB images and depth maps, arising from the four strategies that we compare. For each view, we also report PSNR and SSIM (higher is better). *Closest view* denotes the nearest image in the training set.

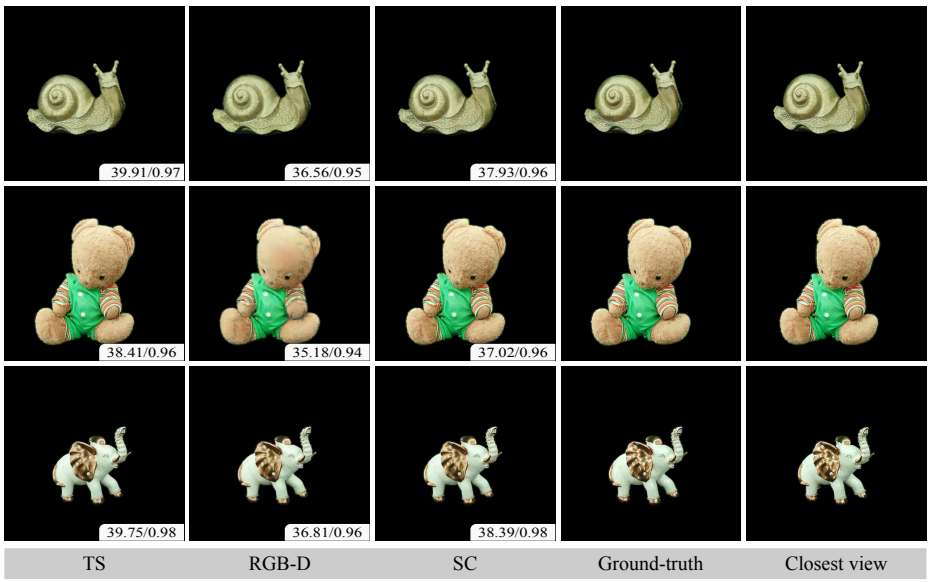

Fig. 8: Reconstructions of a (left-out) RGB image from multi-modal neural scene representations trained on RGB images and depth maps, arising from the four strategies that we compare. FT is left out since its RGB component is similar to TS. For each view, we also report PSNR and SSIM (higher is better). *Closest view* denotes the nearest image in the training set.