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

Mert Özer* Maximilian Weiherer* Martin Hundhausen 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. This work presents four different strategies of how to incorporate thermal imaging into RGB training for Neural Radiance Fields (NeRFs): (1) training from scratch independently on both modalities; (2) pre-training on RGB and fine-tuning on thermal; (3) adding a second branch; and (4) adding a separate component to predict the thermal values. For the evaluation of the proposed strategies, we captured a new multi-view dataset, ThermalMix. Our findings reveal that adding a second branch to NeRF performs best for novel view synthesis on thermal images. We also show that our analysis generalizes to near-infrared images. The ThermalMix dataset is publicly available and can be found at our project page: https://mert-o.github.io/ThermalNeRF/.

1. Introduction

Novel view synthesis involves generating new perspectives from an existing set of images. Historically, this problem has been tackled using conventional techniques, such as structure-from-motion [11], but recently, Neural Radiance Fields (NeRFs) [7] offer a paradigm shift by encapsulating the scene within a continuous radiance field through the adoption of neural networks. On the other hand, multi-modal imaging, characterized by the simultaneous acquisition and processing of multiple data types from different *optical sensors* has shown its significance across myriad applications, ranging from surface reconstruction [1, 4, 16] to medical imaging [13, 17].

In this paper, we conduct a comprehensive evaluation of neural scene representations within a multi-modal context. We propose four different strategies of how to include a second modality into NeRFs: (1) training from scratch, (2) fine-tuning, (3) adding a second branch, and (4) adding a separate component, see Figure 1. We chose thermal imaging as the second modality for this work since we consider modeling thermal images to be one of the hardest (see supp. material). We evaluate the proposed strategies on a newly captured dataset which we name *ThermalMix*. In total, it includes about 360 multi-view RGB and thermal images of six common objects. To summarize, 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 near-infrared images.

A long version of this paper, including supp. material, can be found at https://arxiv.org/abs/2403.11865, published at the ECCV'24 VISION workshop.

2. Related Work

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 [3, 10, 12, 15], RGB-D images [2, 5], or ToF data [6]. Moreover, there are two works that recently tried to build multi-modal NeRFs from RGB and near-infrared images. Based on computed camera poses, [3] 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 [11], X-NeRF [9] *learns* relative poses to the infrared sensor during training, and leverages *Normalized Cross-Device Coordinates* to deal with different camera intrinsics.

*Authors contributed equally to this work.

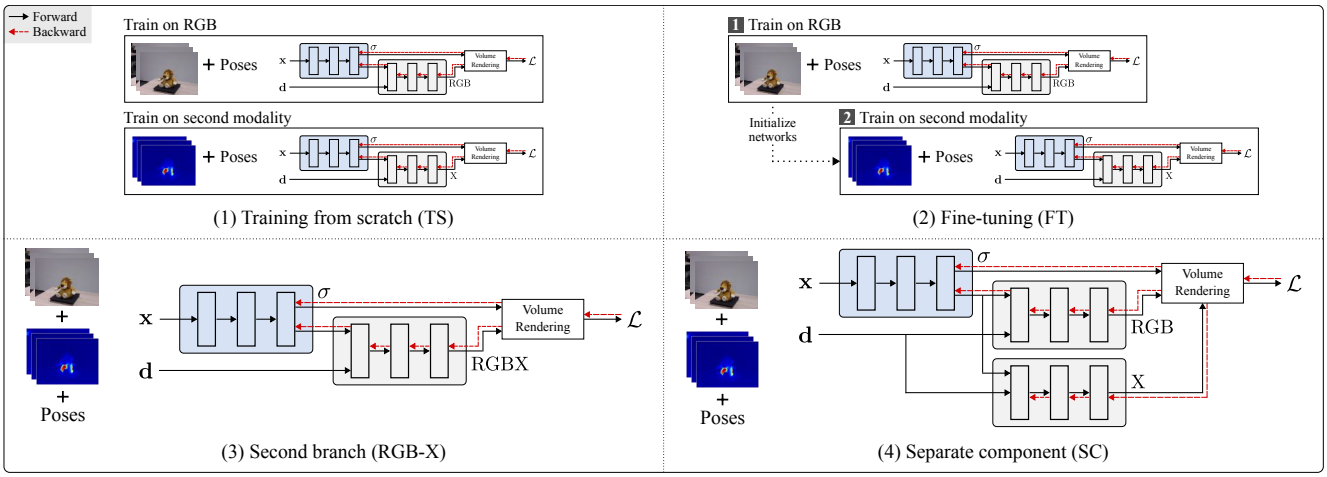

Figure 1. Overview of the four strategies that we compare within this work. In the first strategy (TS), we train a NeRF-like base model (Instant-NGP [8] in our case) from scratch, separately for RGB and the second modality. In the second strategy (FT), we first pre-train our base model on RGB data and 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 (color) 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.

3. Method

We present four strategies to include an additional modality, other than RGB, into neural scene representations, see Figure 1. Throughout this work, we use Instant-NGP [8] as our base model. In brief, Instant-NGP's architecture comprises two fully connected networks for point-wise density and color prediction. The density network generates densities and a geometric descriptor from hash-encoded coordinates and the color network generates view-dependent RGB values from the descriptor and a viewing direction.

In the first strategy (TS), we train our model from scratch, separately for RGB and the second modality. This strategy serves as our baseline. For the training of thermal images, we employ the following NeRF-like loss:

$$\mathcal{L}_t = \sum_{\mathbf{r}\in\mathcal{R}} \left(\hat{t}(\mathbf{r}) - t(\mathbf{r})\right)^2, \tag{1}$$

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. For training on RGB images, we use the standard loss between predicted and true pixel color.

On the other hand, the second strategy (FT) first trains the base model on RGB images and then fine-tunes on images from the second modality. For fine-tuning on thermal images, we apply the same loss as in (1).

The third strategy (RGB-X) utilizes both modalities within a single network by adding a second branch to the color network to predict the values of the second modality. During training, we back-propagate both, RGB *and* predicted values of the second modality through the density network. We use a weighted combination of the thermal and RGB loss functions:

$$\mathcal{L} = \omega_c\mathcal{L}_c + \omega_t\mathcal{L}_t, \tag{2}$$

where we keep $w_c = w_t = 1$ constant (see supp. material for an ablation).

In contrast, our last strategy (SC) adds a separate component to the model that solely predicts values of the second modality but *restricts* back-propagation to the density network during training. We use the same loss as in RGB-X.

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 Figure 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 for both, RGB and infrared sensors. The objects are placed on a table while the camera is moving around the object with a constant distance of about $1\,\mathrm{m}$. Each forward-facing scene contains about 40 images, whereas about 80 images were taken for 360-degree scenes.

A calibration object, whose features are visible in both modalities, is positioned at the center of the scene prior to data capturing, based on which we estimate the relative transformation between the two sensors. Finally, since the distance between the camera and the object is fixed, we compute the camera poses for RGB images using COLMAP

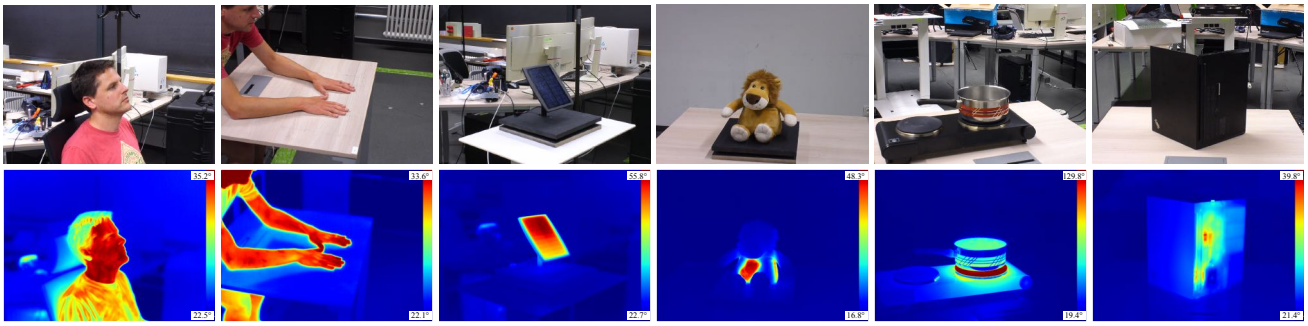

Figure 2. Overview of our newly-captured dataset containing six common objects. FACE, HAND, and PANEL are forward-facing scenes consisting of around 40 images each. LION, PAN, and LAPTOP are 360-degree scenes, where each scene has around 80 images.

	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
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

Table 1. Quantitative results on thermal images, measured using PSNR and SSIM.

[11] and utilize the same poses for aligned thermal images. Please see supp. material for the details of the calibration object and pose estimation.

5. Results

Based on *ThermalMix*, we conducted extensive experiments to compare the proposed strategies on thermal images. Additionally, we used Skoltech3D dataset [14] for the near-infrared (NIR) evaluations.

For pre-processing, RGB and thermal images are normalized to $[0, 1]$. Notably, for thermal images, normalization is performed relative to the *scene's* maximum temperature. Additionally, for FT, we pre-train for 6,000 iterations on RGB and fine-tune for another 4,000 iterations on thermal images, whereas the remaining strategies are trained for 10,000 iterations each. 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 for 10 runs. 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.

Please see long version of the paper for more evaluations and qualitative results.

RGB+Thermal. The results for thermal reconstructions of the forward-facing scenes (FACE, HAND, and PANEL) and 360-degree scenes (LION, PAN, and LAPTOP) are shown in Table 1 and Figure 3. In forward-facing scenes, RGB-t outperforms other methods in terms of PSNR, while TS achieves the highest SSIM due to its reliance on thermal measurements. SC's performance varies due to its lack of thermal density integration, making it scene-dependent. Furthermore, for 360-degree scenes, RGB-t excels in both PSNR and SSIM, whereas TS and FT struggle with PSNR in 360-degree contexts, likely due to static backgrounds (see supp. material). SC ranks consistently second to RGB-t, underscoring the limitations of relying solely on RGB densities in thermal contexts. Please refer to the supp. material for further information.

	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

Table 2. Quantitative results on RGB images, measured using PSNR and SSIM. FT is left out since its RGB component is similar to TS.

Quantitative RGB reconstruction results are shown in Table 2. Since FT's RGB component closely resembles TS, we focus on TS, RGB-t, and SC. TS slightly outperforms other strategies in both PSNR and SSIM. SC delivers superior reconstruction quality 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 RGB images), we find that SC achieves similar reconstruction quality, whereas RGB-t lags slightly behind.

RGB+NIR. Quantitative NIR results are shown in Table 3. For all experiments, we used three forward-facing scenes

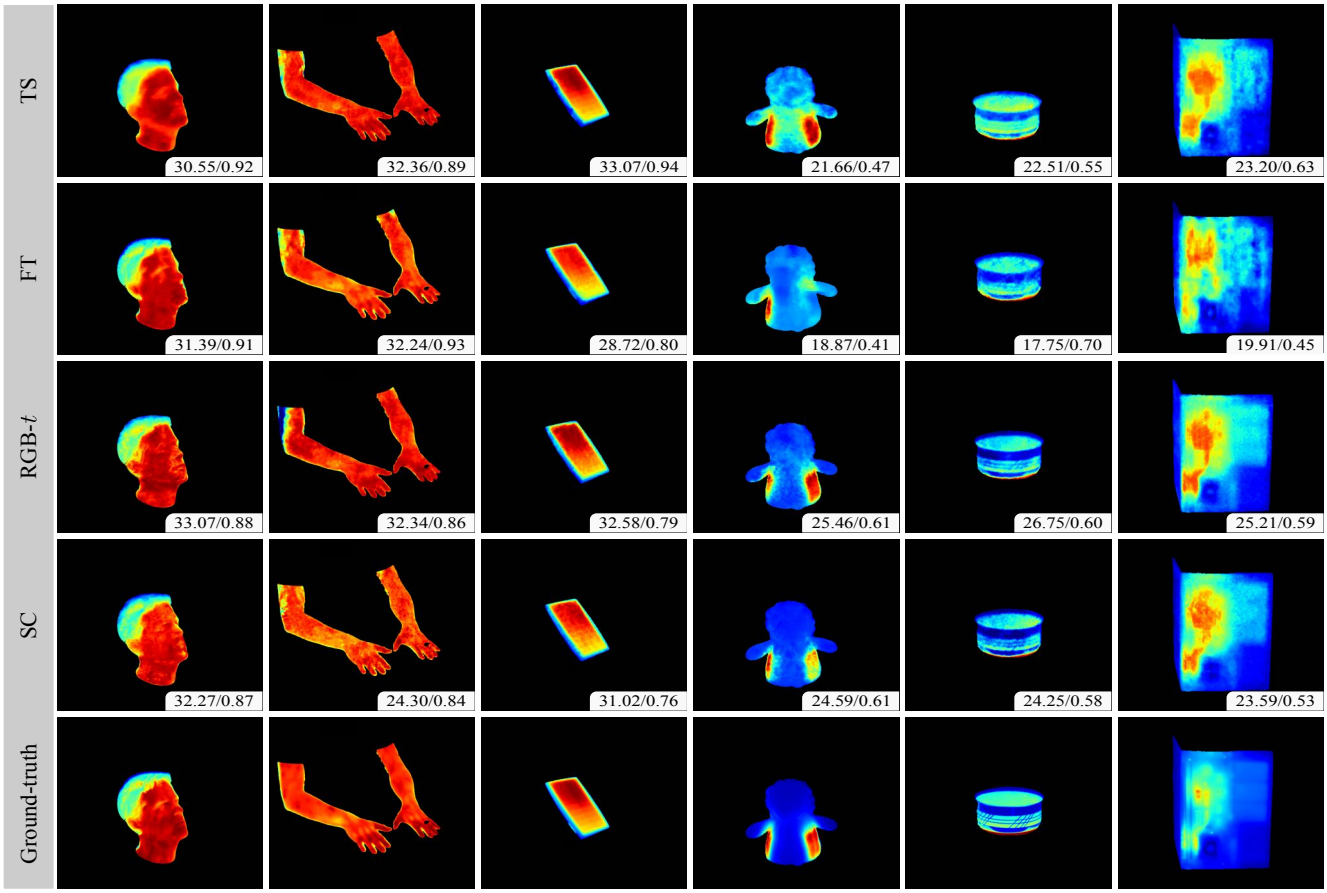

Figure 3. Qualitative results on thermal images. Novel views are rendered using the multi-modal scene representations arising from the four strategies that we compare. For each view, we also report PSNR and SSIM (higher is better).

	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

Table 3. Quantitative results on NIR images.

	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**

Table 4. Quantitative results on RGB images for NIR dataset.

from the multi-sensor dataset proposed in [14]. Considering NIR reconstruction quality, RGB-X (denoted as RGB-NIR) 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 as static and textureless background as thermal images.

Finally, quantitative results on RGB reconstruction quality with the NIR dataset can be found in Table 4. 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. Conclusion

In this paper, we have compared four different strategies of how to incorporate a second modality to NeRFs, other than RGB. We proposed to include 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). The analysis of the four strategies is based on a newly captured publicly available dataset, named *ThermalMix*, which consists of 360 multi-view RGB and thermal images. Our findings indicate that RGB-X stands out for its thermal reconstruction capabilities while also delivering compelling RGB reconstructions. Finally, we also show that our results generalize to NIR images, leading to the conclusion that RGB-X seems to be well-suited for building general multi-modal neural scene representations.

Acknowledgements

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] Koc Ceyhun, Pinarer Ozgun, and Turhan Sultan. 3d mesh model generation from ct and mri data. In *IEEE BigData 2021*, pages 4725–4730, 2021. 1

[2] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, pages 12882–12891, 2021. 1

[3] Zhu Haidong, Sun Yuyin, Liu Chi, Xia Lu, Luo Jiajia, Qiao Nan, Nevatia Ramkant, and Kuo Cheng-Hao. Multimodal neural radiance field. In *ICRA*, pages 9393–9399, 2023. 1

[4] Wenbo Han, Xiaomeng Liu, Shuang Song, and Max Q H Meng. 3d reconstruction of dense model based on the sparse frames using rgbd camera. In *ROBIO*, pages 2726–2731, 2019. 1

[5] Zhu Haoyi. X-nerf: Explicit neural radiance field for multi-scene 360° insufficient rgb-d views. In *WACV*, pages 5766–5775, 2023. 1

[6] Xinyang Liu, Yijin Li, Yanbin Teng, Hujun Bao, Guofeng Zhang, Yinda Zhang, and Zhaopeng Cui. Multi-modal neural radiance field for monocular dense slam with a light-weight tof sensor. In *ICCV*, 2023. 1

[7] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[8] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans Graph*, 41, 2022. 2

[9] Matteo Poggi, Pierluigi Zama Ramirez, F Tosi, Samuele Salti, S Mattoccia, and L D Stefano. Cross-spectral neural radiance fields. In *I3DV*, 2022. 1

[10] Herau Quentin, Piasco Nathan, Bennehar Moussâb, Roldão Luis, Tsishkou D., Migniot C., Vasseur P., and Demonceaux C. Moisst: Multi-modal optimization of implicit scene for spatiotemporal calibration. In *IROS*, 2023. 1

[11] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 1, 3

[12] Tao Tang, Guangrun Wang, Yixing Lao, Peng Chen, Jie Liu, Liang Lin, Kaicheng Yu, and Xiaodan Liang. Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis. *arXiv:2402.17483*, 2024. 1

[13] Rohollah Moosavi Tayebi, Rahmita Wirza, Puteri Suhaiza Sulaiman, Mohd Zamrin Dimon, Fatimah Khalid, Aqeel A Al-Surmi, and Samaneh Mazaheri. 3d multimodal cardiac data reconstruction using angiography and computerized tomographic angiography registration. *J Cardiothorac Surg*, 10, 2015. 1

[14] Oleg Voynov, Gleb Bobrovskikh, Pavel Karpyshev, Saveliy Galochkin, Andrei-Timotei Ardelean, Arseniy Bozhenko, Ekaterina Karmanova, Pavel Kopanev, Yaroslav Labutin-Rymsho, Ruslan Rakhimov, Aleksandr Safin, Valerii Serpiva, Alexey Artemov, Evgeny Burnaev, Dzmitry Tsetserukou, and Denis Zorin. Multi-sensor large-scale dataset for multi-view 3d reconstruction. In *CVPR*, pages 21392–21403, 2023. 3, 4

[15] Qi Zhang, Bo H Wang, Mingli C Yang, and Hang Zou. Mmnerf: Multi-modal and multi-view optimized cross-scene neural radiance fields. *IEEE Access*, 11:27401–27413, 2023. 1

[16] Michael Zollhöfer, Patrick Stotko, Andreas Gorlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum*, 37:625–652, 2018. 1

[17] Örs Petneházy, Shannon Rück, Endre Sós, and László Z Reinitz. 3d reconstruction of the blood supply in an elephant's forefoot using fused ct and mri sequences. *Animals*, 13, 2023. 1