Learning Spectral and Polarimetric Clues for One-to-Multimodal Novel View Synthesis

Federico Lincetto[1], Gianluca Agresti[2], Mattia Rossi[2], Piergiorgio Sartor[2], and Pietro Zanuttigh[1]

[1] MEDIA Lab, University of Padova, Padova, Italy
{federico.lincetto,zanuttigh}@dei.unipd.it
[2] Sony EUISPC, Sony Semiconductor Solutions Europe, Stuttgart, Germany
https://medialab.dei.unipd.it/paper_data/SPoILeR/

Abstract. Neural rendering techniques allow for accurate reconstruction of the geometry and color appearance of 3D scenes. Some methods have extended their use to additional imaging modalities, such as multispectral, infrared, or polarimetric data. However, all of these approaches require expensive sensors and calibrated setups to capture new multimodal frames for each new scene. We propose Spectral and Polarimetric Implicit Learned Representation (SPoILeR), a novel method to obtain multi-view consistent renderings of unconventional modalities for scenes where either only RGB frames or very few of the additional modalities are available. Thanks to a multimodal pre-training phase, the model learns the mutual correlation between different modalities. This step allows predicting accurate renderings of unconventional modalities during a fine-tuning phase supervised only by RGB images. Experimental results show that the approach can accurately render infrared, polarimetric, and multispectral frames for scenes where no input sample captured by these types of sensors is provided.

Keywords: Neural Rendering · Multimodal Data · Modality Conversion

1 Introduction

Neural Radiance Fields (NeRF) [47] represented a turning point for Novel View Synthesis by permitting the rendering of images with a photorealistic appearance and accurate multi-view consistency. NeRF evolved rapidly to solve also related tasks, such as multi-view 3D geometry reconstruction [39, 41, 62, 64, 65, 72], the inverse rendering problem [5, 29, 67, 70, 76], and 3D scene semantic segmentation [5, 11, 12, 31, 43]. Gaussian Splatting (GS) [30] further boosted rendering accuracy at the cost of higher memory requirements and promoted the revision of many tasks previously addressed with NeRF [13, 26, 40, 53, 54, 79]. However, the rapid development of NeRF and GS was possible due to the large amount of available RGB data, which is unquestionably the most popular imaging representation but definitely not the only one. There exists a long-standing interest in exploiting information captured by different imaging sensors to solve complex

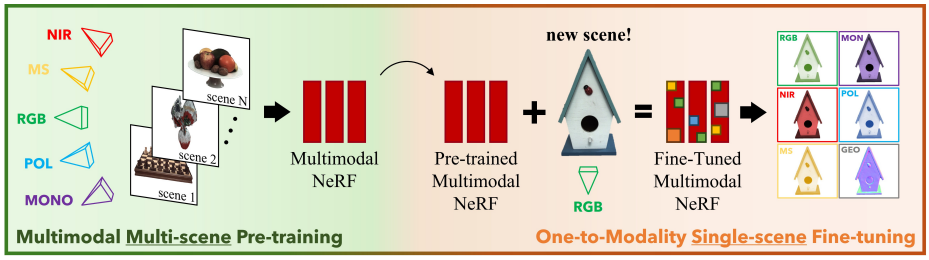

Fig. 1: The proposed approach firstly learns the correlation across multiple modalities on a collection of scenes. Then, a single-scene fine-tuning on RGB data alone produces a model able to render views of arbitrary modalities for the considered scene.

tasks or enhance result quality. For example, there exist Multispectral (MS) sensors that are sensitive to different bands of visible light, Near-Infrared (NIR) sensors that allow sensing radiation with wavelength over 700 nm, and Polarization (Pol) sensors that can measure the polarization of light. Employing MS and NIR data, it is possible to perform tasks such as physics-based rendering [14, 32], image content enhancement [23, 59, 81], and solve ill-posed classification problems [20, 56, 63, 74]. Similarly, polarimetric data provide additional information on surface orientation and material properties, which is particularly helpful for the problems of shape-from-polarization [2,34], material properties estimation [15,35,36], and reflection removal [33,71]. There exist works that exploit data different from RGB for novel view synthesis, 3D reconstruction, or inverse rendering purposes, and that are based on NeRF [15,24,35,37,38,45,50,51,69,73] or GS [10, 22, 44, 58, 61]. However, these methods focus on either a single or a couple of modalities, while just a few works perform novel view synthesis by including a wider set of multiple imaging modalities [21, 32, 42, 52]. One limitation common to all of these works is that they require multiple multi-view consistent images of the same scene, captured by sensors that are more expensive and cumbersome to deploy compared to RGB cameras, thus limiting their practical applicability. Data availability is scarce, and this also prevents the training of large foundation models for these tasks. Moreover, methods that perform conversion from RGB to a different modality [6, 77] lack multi-view consistency.

In this work, we build a multimodal neural representation that overcomes these limitations by learning the mutual correlation that exists between different modalities, given a generic set of scenes for which multimodal information is available. Once done, the learned knowledge can be exploited to render accurate and multi-view consistent novel views of unconventional modalities of a new scene for which either only RGB or a very limited number of other modality frames are available. For this purpose, we introduce several novel contributions:

– A multimodal multi-scene pre-training (PT) procedure able to encode the information captured by different sensors from different scenes in a multimodal latent space. The PT also optimizes decoders for every sensor, capable of mapping the latent space into explicit multimodal radiance information.

- A fine-tuning (FT) pipeline that can operate in a novel scene even if captured with a limited subset of sensors. The FT phase, due to the PT knowledge, estimates the radiance information of missing modalities, thus allowing the model to render photorealistic and multi-view consistent novel views without the need for modality-specific sensor captures of the new scene.
- A regularization procedure consisting of three ad-hoc loss functions that encourage the model to first build a robust and explainable multimodal latent space during the PT and then to preserve the knowledge even with a lack of supervision from other modalities during the FT phase.
- An extensive analysis of the model performance and capabilities with respect to previous multimodal state-of-the-art solutions for spectro-polarimetric modality-to-modality conversion [6, 77] and novel view synthesis tasks [42].

2 Related Work

Multimodal Neural Rendering As anticipated in Sec. 1, there exist methods that tackle the novel view synthesis task by involving multiple imaging modalities. X-NeRF [52] is a precursor in this field: it uses a traditional NeRF-based method to learn the radiance field of forward-facing scenes captured with RGB, NIR, and MS sensors. Recently, a new interpretation based on GS has been released, which reduces the rendering time and improves quality [21]. Another method for multimodal novel view synthesis is NeSpoF [32], which leverages images with a known polarization state for every band of the visible spectrum to learn a spectro-polarimetric radiance field. It models the dependence of polarization on spectral bands for novel view synthesis, but the data acquisition is very demanding in terms of hardware and time. Finally, MultimodalStudio [42] is a recent NeRF-based method that enables novel view synthesis and 3D reconstruction from a wider set of heterogeneous modalities. However, all mentioned methods require capturing a complete set of multimodal images for every new scene to reconstruct. Generally, sensors different from RGB cameras are expensive, and the acquisition procedure may be more complex and time-consuming. With SPoILeR, we address this limitation by reducing the need for multimodal images to a one-off pre-training step that learns the mutual correlations between modalities. This allows one to perform fine-tuning trainings on novel scenes captured by a single modality (*e.g.*, RGB) and still be able to render multimodal frames, thus eliminating the need for expensive multimodal acquisitions.

Modality-to-modality Conversion Several feed-forward methods have been proposed to estimate different modalities from RGB images. Multi-MAE [4] and 4M [3, 48] are popular examples of models that can convert a single RGB frame to a variety of different semantic and geometric modalities. There exist also multi-view methods like RoRE [18] that can achieve consistent novel view synthesis across multiple modalities. However, all these feed-forward models require extensive paired data for training, and thus, due to the lack of spectral and polarimetric data, they are not always viable solutions, particularly when dealing

with custom sensors for which only small datasets exist. On the other hand, StylizedNeRF [27] and StylizedGS [75] enable the possibility of replicating the style of a single conditioning image on novel views. However, when using spectral or polarimetric images as conditioning, they cannot assign the physically correct spectral radiance to the scene objects. Regarding RGB to Hyperspectral (HS) conversion, some methods are based on CNN [57], Vision Transformer (ViT) [16, 68], or they recursively invoke a reconstruction module for a coarse-to-fine estimation [46]. A popular robust baseline for RGB to HS conversion is MST++ [6], the winner of the NTIRE 2022 Spectral Recovery Challenge [1]. It is based on a ViT that employs Spectral-wise Multi-head Self-attention blocks to learn the correlations between RGB and HS spatial and spectral information to perform spectral restoration. Considering polarimetric data, the recently published PolarAnything [77] work converts from RGB to Pol by estimating the Angle of Polarization (AoP) and Degree of Polarization (DoP). For this purpose, the authors fine-tuned Stable Diffusion [55] with synthetic polarimetric data. However, these methods share a common limitation: they predict information that is not multi-view consistent because they consider only a single RGB view at a time. Employing these frameworks to convert a set of multi-view RGB images to MS or Pol data leads to results that exhibit clear inconsistencies between different views. Therefore, in practice, they are not reliable enough to support multimodal novel view synthesis. In contrast, SPoILeR can produce multimodal renderings that are multi-view consistent by construction, given that it optimizes a single 3D representation. In Sec. 4.5, we validate these claims by comparing SPoILeR with state-of-the-art modality conversion strategies such as MST++ and PolarAnything.

Cross-scene Novel View Synthesis Vanilla NeRF and GS are generally optimized for single scenes. However, there exist methods that propose training on several scenes to enable zero-shot novel view synthesis of new scenes [8, 25, 28, 60, 66, 78, 80, 82]. All these methods rely on CNN or ViT models which are known to be data hungry for training. On the other hand, Dictionary Fields (DiF) [9] proposes an implicit representation decoupled into coefficient and basis modules, suitable for addressing the few-shot novel view rendering task as well. The key advantage of this approach is that its architecture explicitly induces a separation of cross-scene and scene-specific information, without requiring large amounts of data to train. In SPoILeR, we reinterpret this concept to promote the learning of shared multimodal knowledge. Unlike DiF, our FT step is dense in terms of multi-view coverage; instead, we leverage the basis and coefficient architecture to address the much sparser modality coverage of the FT with respect to the PT. Moreover, unlike DiF, we explicitly define and regularize a multimodal latent space that learns to model the mutual correlations between modalities.

3 Method

In this section, we present SPoILeR, a multimodal NeRF-based approach with generalization capabilities across different scenes and imaging modalities. As

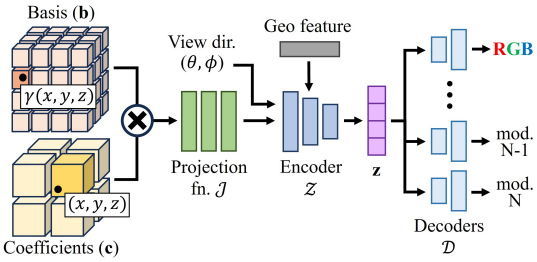

Fig. 2: Radiance module architecture scheme: features from basis and coefficients are combined and decoded into different radiance modalities.

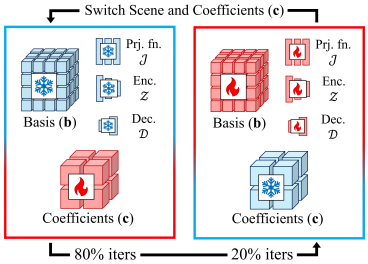

Fig. 3: PT routine visualization: scene-specific and shared modules are alternatively optimized.

shown in Fig. 1, we assume to have a generic set of training scenes with multi-view images captured by different sensors. The goal is to learn a multimodal radiance field of a new scene for which only RGB or very few modalities are available. This allows for the recovery of an accurate and multi-view consistent multimodal representation that enables the photorealistic novel view synthesis of any modality included in the set of sensors. For this purpose, we first propose to learn multimodal knowledge during a PT phase from a set of scenes. Then, a subsequent FT phase adapts only part of the model to new scenes.

3.1 Model Architecture

The SPoILeR NeRF-based architecture is built on top of MultimodalStudio [42], a state-of-the-art approach for multimodal neural rendering. As standard NeRF models, MultimodalStudio Framework (MMS-FW) is scene-specific, thus it cannot transfer knowledge between different scenes. In contrast, our goal is to build a strategy that infers missing modalities even when the supervision is limited to a single modality. Therefore, we employ an architecture that allows us to store and preserve the learned multimodal priors, but is also flexible enough to estimate the geometry and spectral properties of specific scenes. We take inspiration from the DiF architecture [9] and substitute the multi-resolution hash-grids adopted by MMS-FW with two sets of learnable feature grids that act as basis and coefficient fields, respectively, as shown in Fig. 2. Our model trains on frames of N different modalities, whose channel sum is $C = \sum_{l=1}^{N} C_l$. Following the notation of [9], we estimate a radiance signal $\mathbf{s} : \mathbb{R}^3 \times \mathbb{S}^2 \rightarrow \mathbb{R}^C$, with $\mathbf{r}(\theta, \phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ and $\mathbb{S}^2 = \{\mathbf{r}(\theta, \phi) | 0 \leq \theta < \pi, 0 \leq \phi < 2\pi\}$, defined as follows:

$$\mathbf{s}(\mathbf{x}, \mathbf{v}) = \mathcal{P}(\mathbf{c}(x)^\top \mathbf{b}(\gamma(\mathbf{x})), \mathbf{v}), \tag{1}$$

where $\mathbf{x} = (x, y, z)$ is the 3D position, $\mathbf{v} = \mathbf{r}(\theta, \phi) \in \mathbb{S}^2$ the camera viewing direction, $\mathbf{c} : \mathbb{R}^3 \rightarrow \mathbb{R}^D$ and $\mathbf{b} : \mathbb{R}^{3\times D} \rightarrow \mathbb{R}^D$ the coefficient and basis fields, and D the learnable feature dimension. The function $\gamma : \mathbb{R}^3 \rightarrow \mathbb{R}^{3\times D}$ is a periodic coordinate mapping and $\mathcal{P} : \mathbb{R}^D \rightarrow \mathbb{R}^C$ is a function that predicts radiance values. The basis field $\mathbf{b}$ and the function $\mathcal{P}$ are common to all scenes and responsible for learning the radiance correlation shared between different modalities.

The coefficient field $\mathbf{c}$ is optimized specifically for each scene to account for the scene-specific properties, such as textures and reflections. The periodic mapping function γ, which is a multi-frequency sawtooth function, allows us to apply the same basis to different coordinates, thus promoting their generalization. In our formulation, $\mathcal{P}$ is the combination of multiple functions: a projection function $\mathcal{J} : \mathbb{R}^D \rightarrow \mathbb{R}^D$, a latent encoder $\mathcal{Z} : \mathbb{R}^D \rightarrow \mathbb{R}^Z$, with Z as the latent space dimensionality, and a modality decoder $\mathcal{D}_l : \mathbb{R}^Z \rightarrow \mathbb{R}^{C_l}$ specific to each modality. The projection function $\mathcal{J}$ is responsible for manipulating the product of basis and coefficient features; the latent encoder $\mathcal{Z}$ receives the projection function output, concatenates it with the viewing direction $\mathbf{v}$ and a geometry feature $\mathbf{g}$ from the geometry module and returns the latent vector $\mathbf{z}$; the decoders $\mathcal{D}_l$ decode $\mathbf{z}$ into the desired modality channels. Overall, $\mathcal{P}$ is defined as follows:

$$\mathcal{P}(\mathbf{f}, \mathbf{v}, \mathbf{g}, l) = \mathcal{D}_l(\mathcal{Z}(\mathbf{v}, \mathbf{g}, \mathcal{J}(\mathbf{f}))), \tag{2}$$

with $\mathbf{f} = \mathbf{c}^\top \mathbf{b}$. MMS-FW estimates geometry and radiance with two dedicated modules. The per-scene geometry module is based on Signed Distance Field (SDF) [17,49] and returns a density value and a geometry feature $\mathbf{g}$. We maintain the geometry module unaltered. All the functions forming $\mathcal{P}$ are implemented as shallow Multi-Layer Perceptron (MLP) in order to place most of the model capacity in the coefficient and basis fields. This is because we need a latent space that is not too compressed, as it must be easily and consistently estimable even when only a small subset of supervision modalities is available. One more reason is that we aim to split the information into either general or scene-specific: general information must be stored by the basis field to be available in any scenario, while scene-specific information must be stored only in the coefficient field, as it must be completely switched when the scene changes.

The model is trained by minimizing an objective function that includes a photometric loss averaged between modalities $\mathcal{L}_{mod}$, the eikonal loss $\mathcal{L}_{eik}$, and curvature loss $\mathcal{L}_{curv}$ for geometry regularization [19,39], and three new regularization losses ($\mathcal{L}_{lsg}$, $\mathcal{L}_{inv}$, and $\mathcal{L}_{m2l}$, detailed in Sec. 3.4) that promote a more robust and explainable latent space. The full loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{mod} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{curv}\mathcal{L}_{curv} + \lambda_{lsg}\mathcal{L}_{lsg} + \lambda_{inv}\mathcal{L}_{inv} + \lambda_{m2l}\mathcal{L}_{m2l}, \tag{3}$$

where the λ terms are the weighting coefficients (see Sec. 4.1).

3.2 Pre-training Pipeline: Shared Modules Optimization

As anticipated previously, the model training is organized in two distinct phases: a pre-training phase and a fine-tuning phase. The PT phase is meant to let the model learn a basis representation for the radiance, able to generalize well between different scenes, and the correlation that exists between different imaging modalities and different scene properties. For this purpose, the model is trained to estimate the radiance field of several scenes captured by the complete set of multimodal sensors. At the same time, the unique radiance and geometry properties of every scene are estimated by independent scene-specific coefficient fields

and modules, respectively. All other modules, namely the projection function, the latent encoder, and the modality decoders, are shared between scenes. To promote an effective separation of shared and specific information, the training freezes shared and scene-specific modules alternately, with a fixed schedule, as shown in Fig. 3. When loading a scene, the corresponding coefficient module is loaded too, and the model is trained for B iterations. The scene-specific modules (namely, coefficients and geometry module) and the shared modules are trained for 80% and 20% of the B iterations, respectively. Then, a new scene is loaded, coefficients are swapped, and the process is repeated. It is important to carefully consider the role of the geometry feature $\mathbf{g}$ produced by the geometry model and passed as input to the latent encoder $\mathcal{Z}$, part of the radiance module. Previous works introduced this feature to improve geometry estimation [72]. However, in our setting, there exists the possibility that part of the radiance information, that should be stored in the shared modules to remain available during the FT process, is instead stored in the geometry feature, which is scene-specific. We found that the best solution to prevent this is to employ a dropout of its gradient. During the PT, only the gradient of one random modality per iteration can back-propagate through the geometry feature. We observe that this solution produces good geometry estimation while preserving the shared radiance information for the FT step.

3.3 Fine-tuning Pipeline: Coefficient Field Optimization

In the FT phase, we get a new scene, not available during the PT phase, captured only by a subset of the imaging modalities. Leveraging the pre-trained model, the goal is to recover photorealistic renderings of modalities not available in the FT phase but available in the PT. For this reason, we need to exploit the knowledge stored in the pre-trained shared modules but, given that the multimodal supervision signal is now limited, we also need to preserve the learned correlations between modalities as much as possible to avoid catastrophic forgetting. At the same time, the model must be flexible to learn the new scene-specific geometry and radiance. Therefore, during FT we freeze the shared modules and optimize only the scene-specific radiance coefficients and geometry field. As a result, the much smaller number of learnable parameters than during PT leads to a much shorter training time than re-training a complete model from scratch on the new scene. The training objective for the FT matches the PT one, except for $\mathcal{L}_{lsg}$ which is disabled since the latent space geometry must be preserved at this stage.

3.4 Latent Regularization Strategies

Despite the proposed architectural design and training schedule, as experienced through empirical evaluation, if the model is trained only on a subset of the modalities, the optimization is prone to diverge for the modalities not optimized. For this reason, we introduce 3 latent regularization losses: 1) the latent space geometry loss, 2) the inverse function loss, and 3) the modality-to-luma loss.

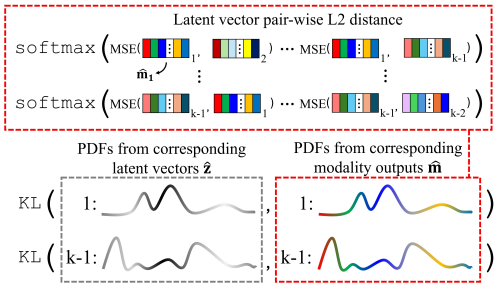

Fig. 4: The latent space geometry loss encourages the model to produce latents whose pairwise distance distribution matches the radiance pair-wise distance distribution.

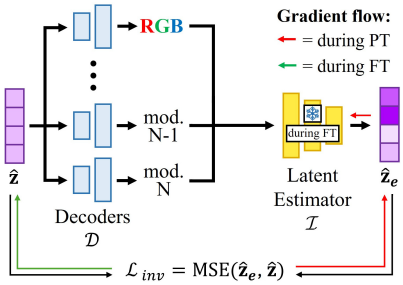

Fig. 5: Scheme of the inverse function module $\mathcal{I}$. It predicts the estimated latent vector $\hat{\mathbf{z}}_\mathbf{e}$ from multimodal radiance values $\{\hat{\mathbf{m}}_1, ..., \hat{\mathbf{m}}_k\}$.

Latent Space Geometry Loss This regularization loss aims to promote the latent space explainability. Considering that the multimodal supervision is strongly limited in the FT, it is reasonable to observe the model estimating suboptimal latents. We observed that the decoding procedure is sensitive to small latent perturbations because similar latents are sometimes decoded into very different radiance values. This issue reduces the capability of the FT phase to obtain accurate renderings of the missing modalities. To overcome this issue, inspired by the embedding loss proposed in [7], we introduce the latent space geometry loss $\mathcal{L}_{lsg}$. The goal is to build a latent space where the distances between latent representations reflect the distances between the corresponding decoded modalities, as shown in Fig. 4. Unlike in [7] where the aim is to align two latent spaces, here we treat the modality space as if it were the latent space whose distribution has to be matched. Let's consider a set of k latent vectors $\{\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_k\}$ and the corresponding set of decoded modality channels $\{\hat{\mathbf{m}}_1, ..., \hat{\mathbf{m}}_k\}$, with $\hat{\mathbf{m}}_k$ being the concatenation of all N modality channels such as $\hat{\mathbf{m}}_k = (C_{k1}, ..., C_{kN})$. We compute the pairwise Euclidean distance between all pairs of latents and modality channels independently, thus building two $(k-1) \times (k-1)$ matrices, D_z and D_m. Then, we compute the softmax row-wise in order to extract two PDFs for each pair of latent vectors and the corresponding modality channels, and we minimize the difference between the two by applying the Kullback–Leibler (KL) divergence, as shown in Fig. 4. Therefore, $\mathcal{L}_{lsg}$ is defined as follows:

$$\mathcal{L}_{lsg} = \texttt{KL}(\texttt{softmax}_{row}(-D_z), \texttt{softmax}(-D_m)). \tag{4}$$

Note that the "hat" notation above disambiguates the latent vectors $\mathbf{z}$, predicted for each ray sample, from the latent vector $\hat{\mathbf{z}}$, which is the result of the integration of latent vectors along each ray. This solution is adopted to reduce the computational demand that otherwise would be unaffordable. Finally, we disable this loss during the FT, as the latent space geometry is already formed.

Inverse Function Loss The second regularization loss that we propose aims to promote the latent space consistency between the PT and the FT phases.

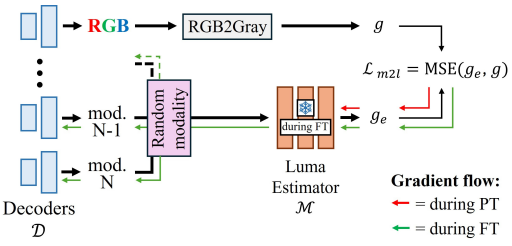

Fig. 6: Scheme of the modality-to-luma module $\mathcal{M}$. It encourages consistency between the reference RGB luminance g and the estimated luminance g_e predicted from one random modality.

As shown in Fig. 5, we train a MLP module that, given the modality radiance values decoded from a latent vector, estimates the latent vector itself. We refer to it as inverse function because it performs the opposite operation done by the latent space decoders. Specifically, the inverse function $\mathcal{I} : \mathbb{R}^C \rightarrow \mathbb{R}^Z$ is defined as $\hat{\mathbf{z}}_e = \mathcal{I}_\Theta(\hat{\mathbf{m}})$. This module is trained during the PT phase to learn the mapping between modality space and latent space. In this phase, we prevent the gradient from back-propagating through the whole model to avoid influencing the latent space formation. Then, in the FT phase, the inverse function module is frozen, and the estimated latent vector acts as a reference for the latent vector encoded by $\mathcal{Z}$. We compute the mean squared error between the actual and estimated latent space and back-propagate the gradient through the whole model to guide the optimization of the coefficients $\mathbf{c}$. Therefore, the loss is defined as $\mathcal{L}_{inv} = \texttt{MSE}(\hat{\mathbf{z}}_e, \hat{\mathbf{z}})$.

The purpose of this loss is to encourage the FT latent space to be coherent with the latent space learned during PT; otherwise, it is likely that the latent space diverges and, consequently, the renderings of the estimated modalities.

Modality-to-luma Loss The last regularization loss is the modality-to-luma loss function. As shown in Fig. 6, we train a shallow MLP that estimates the function $\mathcal{M} : \mathbb{R}^{M_l} \rightarrow \mathbb{R}$, where M_l refers to the number of channels of modality l. Given all channels of a selected modality $\hat{\mathbf{m}}_l$, this function returns a grayscale value g_e, defined as $g_e = \mathcal{M}_\Theta(\hat{\mathbf{m}}_l, l)$, that matches the luminance g computed from the RGB channels. Therefore, the modality-to-luma loss is computed as $\mathcal{L}_{m2l} = \texttt{MSE}(g_e, g)$. In this case, as before, we train the module during the PT by detaching the gradient, then use it frozen during FT. In our case, all modalities integrate light from the visible spectrum; therefore, they are indeed correlated. However, some of them (*e.g.*, near-infrared, mono, and polarization) also integrate part of the infrared spectrum. Because of this, we observed that the model tends to ignore their correlation with other modalities and decouple their representations into more orthogonal ones. The modality-to-luma loss encourages the model to preserve the modality correlation during FT, when complete multimodal supervision is missing. However, its effectiveness is subject to the chosen modality set because it is not always the case that their values are correlated with RGB radiance (*e.g.*, for thermal images). Considering that our modality set satisfies this constraint, we believe that its use is motivated in our scenario. Finally, it is worth noting that the inverse function and the modality-to-luma losses act as complementary regularizations: the inverse function loss promotes the preservation of latent space, particularly for well correlated modalities; the modality-to-luma loss acts as guidance for the less correlated ones.

4 Experimental Evaluation

In this section, we present a set of experiments that demonstrate the effectiveness of SPoILeR. We recall that the main goal of this work is to build a method that can learn the mutual correlation between different imaging modalities in order to enable the generation of photorealistic renderings of any of them, even when the scene is captured by a single or a limited subset of sensors. For this reason, we focus on the FT results while including most of the PT results in the supplementary material. In the next sections, we show that SPoILeR can produce accurate renderings of modalities with limited or no supervision. We also include comparisons with single-view RGB-to-modality methods such as MST++ [6] and PolarAnything [77], proving that SPoILeR can outperform them in terms of multi-view radiance consistency.

4.1 Implementation Details

We recall that the model is trained in 2 stages: a multi-scene pre-training followed by a scene-specific fine tuning. The PT runs for 1M iterations on a single NVIDIA RTX 6000 Ada. In the PT, the parameter B (Sec. 3.2) is set to 15 iterations, while λ_{eik} and λ_{curv} are set at 0.1 and 0.0005, respectively, as suggested in [39]. The regularization loss weights are set to $\lambda_{lsg} = 0.1$, $\lambda_{inv} = 1.0$, and $\lambda_{m2l} = 1.0$. The scene-specific FT runs for only $30k$ iterations using $\lambda_{inv} = 0.01$ and $\lambda_{m2l} = 0.02$, while the latent space geometry loss is disabled in this phase.

To evaluate the results, we consider MMS-FW as a reference and compare with its metrics. However, it is important to mention that defining a fair comparison is not trivial. MMS-FW is trained from scratch on a single scene, optimized for 100k iterations, and its radiance module counts ∼12M parameters. On the other hand, our FT model is optimized for 30k iterations and its radiance module has ∼4.1M parameters, but only ∼7% of them are trainable (∼300k parameters). The other ∼3.8M are frozen because optimized only during the PT for ∼200k iterations on several different scenes. However, the PT step must be completed just once, thus it is reasonable not to consider it when evaluating the cost of the per-scene optimization. The choice to use a smaller model with respect to [42] is motivated by the fact that the PT phase has high memory requirements in order to handle several scenes at the same time. Given the differences in architectures and training pipelines, it is impossible to align the model capacities and the number of training iterations. For these reasons, we opt to preserve the original architecture of MMS-FW, but it is worth noting that SPoILeR has 3 times smaller capacity and the fine-tuning time is 4 times shorter.

4.2 Datasets and Metrics

The experiments are conducted on MMS-DATA [42], a multi-view multimodal dataset encompassing 5 different modalities: *i.e.*, RGB, NIR, Mono, Pol, and MS images. It includes 32 object-centric scenes. The PT is performed on 27 scenes, and the FT on the remaining 5 scenes. Given that SPoILeR is built on

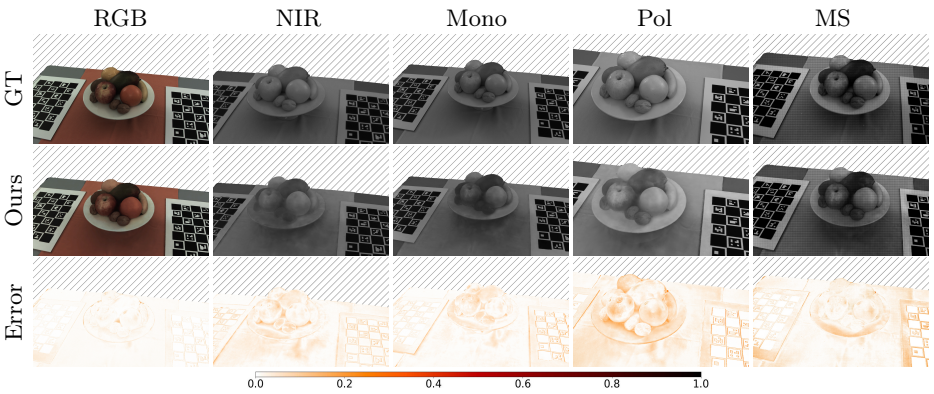

Fig. 7: Qualitative samples from the "Fruits" scene FT performed with only RGB data

MultimodalStudio [42], we follow the same strategy of training our model with MMS-DATA raw mosaicked frames and computing the metrics only on masked foreground regions. To evaluate the quality of all rendering results, we use Peak Signal-to-Noise Ratio (PSNR) as the main metric. In addition, we also show Structural Similarity Index Measure (SSIM) for single channel or demosaicked images, since its use is not appropriate with mosaicked data [42].

4.3 Multi-scene Pre-training and Single-scene Fine-tuning

As described in Sec. 3, SPoILeR is trained in two stages: firstly, a multi-scene PT phase and then a single-scene FT phase. The output of the FT stage is the final outcome that we discuss in this section. The main purpose of the PT is to learn the scene-independent mutual correlations between modalities to effectively support the lack of multimodal data during the FT, rather than accurately reconstructing the pre-training scenes. The PT rendering accuracy, reported in the supplementary material, is slightly lower than FT results, and this is reasonable as, unlike single-scene methods, its focus is more on generalization. Considering the FT metrics shown in Tab. 1, it is possible to compare the results achieved by SPoILeR and MMS-FW. When supervised by a single modality, namely RGB, it is interesting to note that SPoILeR achieves a higher PSNR than MMS-FW. This result highlights that the PT knowledge, built across several modalities and scenes, helps to improve the quality also of supervised modalities with respect to a model trained from scratch on a single scene. More importantly, it also allows rendering additional modalities not part of the FT training images, which is impossible for MMS-FW and for standard novel view synthesis approaches.

In Tab. 1, we also show the results achieved by SPoILeR and MMS-FW when supervised by all modalities. These results serve as upper bounds for our FT model with single-modality supervision. As expected, MMS-FW outperforms SPoILeR, due to its larger capacity and the training on a single scene from scratch. We observe that the improvement driven by these factors can be quantified in $\sim$2 dB of PSNR, while SSIM is perfectly comparable. In Fig. 7, we show some qualitative results; additional results are in the supplementary material.

Table 1: Results of SPoILeR (Ours) FT and of MMS-FW [42], averaged on the five FT scenes. Standard case: both models are trained only with RGB frames. Upper bounds: both models are supervised with frames of all modalities.

		Train Mod.	Test Mod.	PSNR↑	SSIM↑
Standard case	Ours	RGB	RGB	30.07	-
			Mono	25.78	0.88
			NIR	26.55	0.87
			Pol	24.25	-
			MS	25.45	-
	MMS	RGB	RGB	29.53	-
Upper bounds	Ours	ALL	RGB	30.54	-
			Mono	29.21	0.93
			NIR	32.12	0.92
			Pol	29.32	-
			MS	28.46	-
	MMS	ALL	RGB	32.77	-
			Mono	32.98	0.94
			NIR	34.25	0.93
			Pol	30.66	-
			MS	31.42	-

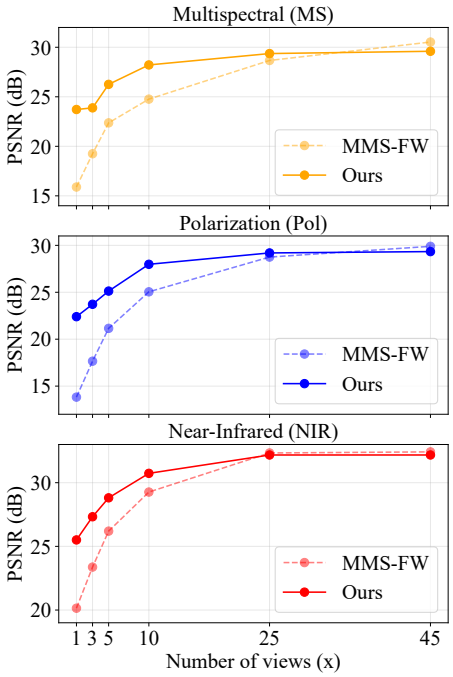

Fig. 8: Tests with an unbalanced combination of modalities. X axis refers to the number of additional modality (MS, Pol, or NIR) frames. Comparison between SPoILeR (Ours) and MMS-FW.

4.4 Unbalanced Combination of Modalities

In this section, we investigate the results achieved by MMS-FW when trained with an unbalanced combination of modalities. This test is introduced by [42] and involves scenarios where many frames of one modality but only a few of a second one are available, thus the model receives stronger supervision for the first modality and limited guidance for the second. This scenario represents situations where, due to deployment or cost constraints, it is impossible to capture the same amount of frames for every modality. At the same time, even when it is possible, it demonstrates that having full scene coverage for all modalities is not mandatory. To prove this claim, we perform some tests involving RGB-MS, RGB-Pol, and RGB-NIR frames. Following the idea that RGB cameras are cheaper and more widely accessible, all RGB frames (45) are always available, while the second-modality frames are reduced to 1, 3, 5, 10, and 25. In Fig. 8, we show the PSNR trend of the rendered second-modality frames as a function of the number of available frames. We observe that SPoILeR clearly outperforms MMS-FW, particularly in scenarios with few second-modality frames available. When only one MS, Pol, or NIR frame is available, SPoILeR achieves higher PSNR, with gains of ∼8, ∼7, and ∼6 dB, respectively. With 10 second-modality

Table 2: Comparison between SPoILeR and MMS-FW on MS frames. MMS-FW is trained with MS frames estimated from RGB using MST++, denoted as MS*.

	Train	Test	PSNR↑
MMS	MS*	MS	21.99
Ours	RGB	MS	**25.45**

Table 3: Comparison between SPoILeR and MMS-FW on Pol rendering quality. MMS-FW is trained with Pol frames estimated from RGB by PolarAnything, denoted as Pol*. AoP MAngE ranges in $[0°, 90°]$, DoP MAbsE ranges in $[0, 1]$.

	Train	Test	MAngE(°)↓	MAbsE↓
MMS	Pol*	Pol	44.13	0.089
Ours	RGB	Pol	**30.05**	**0.045**

frames, SPoILeR still gains ∼2.5 dB on average, and MMS-FW achieves similar performance as our approach only with 25 frames. These results confirm the positive effect of the PT phase: by leveraging this knowledge, the FT can more easily infer the second-modality radiance even with very few available frames. In these tests, for the sake of fair comparison, both SPoILeR fine-tuning and MMS-FW are trained for 100k iterations. Note that the only other public multi-view multimodal dataset that could be used, *i.e.*, X-NeRF, is not very informative since all scenes are forward-facing and even a single view encompasses almost the whole scene, thus the benefits of the PT are not appreciable. However, we replicated the test on X-NeRF: as expected, SPoILeR performs on par with MMS-FW. We include results on X-NeRF data in the supplementary material.

4.5 Comparisons with Third-party Modality Conversion Methods

In literature, there exist many solutions to predict multimodal data starting from RGB images, thus a viable strategy for multimodal novel view synthesis is to use them together with a standard NeRF/GS model. As discussed in Sec. 2, it is possible to employ methods such as MST++ [6] and PolarAnything [77] that infer MS and Pol information, respectively, from RGB frames with a feed-forward single shot approach. However, these methods work on a single view and could produce results that lack multi-view consistency. To prove this claim, a possible test is to convert single images from RGB to other modalities using MST++ or PolarAnything and then feed the output to a NeRF model, such as MMS-FW. We performed some additional tests to compare this approach with SPoILeR on MS or Pol frames. We start by using MST++ to estimate MS data. Then, we train MMS-FW on these scenes with only the converted MS images and render the evaluation frames that are then compared with the GT. This allows us to compute the metrics between the renderings and the GT frames. Results are in Tab. 2: we observe that SPoILeR outperforms MMS-FW coupled with MST++ in terms of PSNR by 3.46 dB. This happens because, as expected, MST++ does not produce multi-view radiance consistent frames. Therefore, inaccurate information is propagated and averaged through the 3D volume by the NeRF, reducing the accuracy of the final rendering. We perform analogous tests for Pol frames converted from RGB with PolarAnything. As proposed by [77], the polarization can be evaluated by considering the AoP and

Table 4: Ablation study PSNR results. Between brackets: difference with respect to the full model. Results of FT on RGB frames only.

Mod.	Ours	w/o $\mathcal{L}_{lsg}$	w/o $\mathcal{L}_{inv}$	w/o $\mathcal{L}_{m2l}$
RGB	**30.07**	29.90(-0.17)	29.96(-0.11)	29.95(-0.12)
Mono	**25.78**	25.44(-0.34)	25.54(-0.25)	21.04(-4.38)
NIR	**26.55**	26.30(-0.24)	25.96(-0.58)	22.91(-3.64)
Pol	**24.25**	23.08(-1.18)	24.01(-0.24)	23.73(-0.52)
MS	**25.45**	23.14(-2.31)	25.28(-0.17)	25.36(-0.09)

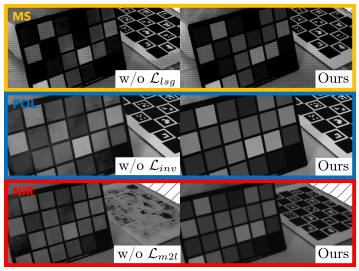

Fig. 9: MS, Pol, and NIR renderings with different loss ablations

DoP in terms of mean angle error (MAngE) and mean absolute error (MAbsE). In Tab. 3 we compare the results achieved by SPoILeR fine-tuning with those of MMS-FW using PolarAnything data. Our model outperforms the competitor in terms of both MAngE and MAbsE, by 14.08° and 0.044, respectively. In this case, PolarAnything cannot generate multi-view consistent Pol frames because it was trained on single-view images. Therefore, combining the contributions of frames from different views leads to very inaccurate results when training a NeRF model. In contrast, SPoILeR relies on the PT knowledge, built across multiple different 3D scenes, thus it is more robust when predicting different views of one missing modality. In conclusion, SPoILeR confirms its reliability and proves to be the current state-of-the-art in terms of multi-view consistency for the task of modality-to-modality conversion. Details about the conversion procedure and qualitative results by MST++ and PolarAnything are in the supplementary material.

4.6 Ablation Study

Finally, we perform an ablation study on the model components. In Tab. 4, we compare the results of the FT with the results of the model when ablating each of the loss functions introduced in Sec. 3.4. Firstly, we observe that every ablation leads to worse results in terms of PSNR. Looking at the results, the modality-to-luma $\mathcal{L}_{m2l}$ loss appears to be the most impactful, particularly for Mono and NIR data, even though Pol is also slightly affected. One likely reason is that these three sensors also measure part of the infrared spectrum, albeit with different sensitivities. Therefore, the luminance information carried by these frames may diverge from what measured by the RGB or MS sensor, thus the model is likely to decouple the information into more orthogonal latent representations. Therefore, when supervising only with RGB frames, the model struggles more to recover the correct Mono, NIR, and Pol information. In contrast, by applying the $\mathcal{L}_{m2l}$ loss, we make the correlation between all modalities more explicit, thus encouraging the model to learn a shared mutual representation.

Similar considerations can be made for the inverse function loss $\mathcal{L}_{inv}$. As mentioned in Sec. 3.4, the purpose of the loss is complementary to the previous one. The inverse function encourages the model to maintain the latent space geome-

try learned during the PT stable. In contrast to the $\mathcal{L}_{m2l}$ that pulls modalities' radiance towards RGB luma, the $\mathcal{L}_{inv}$ aims at preserving the mutual correlations learned over several scenes and when all the modalities are supervised. In other words, it acts as an antagonist to the $\mathcal{L}_{m2l}$ to promote latent space stability.

Finally, we observe that the lack of latent space geometry loss $\mathcal{L}_{lsg}$ produces worse results for Pol and MS renderings, while the other modalities are less affected. The purpose of this loss is to make the latent space more explainable, as it encourages the model to have a latent space that shares consistent distance measures with the corresponding explicit radiance space. In other words, the decoded radiance values are less sensitive to small shifts in their latent representation. MS and Pol are the modalities with the highest information content, and the channels of each of them are highly correlated. Therefore, it is sufficient that only part of the MS or Pol channels is well correlated with the supervised modality to predict a latent vector that, even if sub-optimal, thanks to the regularized latent space geometry, is decoded to an approximately correct radiance.

5 Limitations

SPoILeR demonstrates the capability of rendering accurate and multi-view consistent spectral and polarimetric frames from a subset of modalities. However, its design inherently carries some limitations worth mentioning. For example, the estimation of spectral properties is reliable only for materials similar to those observed during the PT. Furthermore, a new PT stage must be performed to predict new modalities. Finally, for consistent predictions, the PT images must be captured by devices with fixed exposure and white balancing, and with controlled illumination. Future work will focus on relaxing these constraints by making the model robust to variations of the illumination spectrum or of the sensor properties.

6 Conclusions

In this paper, we propose a novel multimodal neural rendering method that can decouple shared multimodal information from scene-specific content. This allows learning mutual correlations among different modalities from a set of scenes and then exploiting this information to render novel multimodal views of new scenes for which only RGB information is available. The experimental results demonstrate the effectiveness of our method and show that it outperforms both multimodal NeRFs and two-stage approaches that employ third-party methods for modality conversion.

Acknowledgment

This collaborative work is funded by Sony Europe Limited. The work of P. Zanuttigh is also partially supported by the Italian Ministry of University and Research (MUR) under the PRIN 2022 SEQUOIA project (code 2022KCKYA2).

References

1. Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., et al.: Ntire 2022 spectral recovery challenge and data set. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop. pp. 862–880. IEEE (2022). https://doi.org/10.1109/CVPRW56347.2022.00102
2. Ba, Y., Gilbert, A., Wang, F., Yang, J., Chen, R., Wang, Y., Yan, L., Shi, B., Kadambi, A.: Deep shape from polarization. In: European Conference on Computer Vision. pp. 554–571. Springer (2020). https://doi.org/10.1007/978-3-030-58586-0_33
3. Bachmann, R., Kar, O.F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., Zamir, A.: 4m-21: An any-to-any vision model for tens of tasks and modalities. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 61872–61911. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-1977
4. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision. pp. 348–367. Springer (2022). https://doi.org/10.1007/978-3-031-19836-6_20
5. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decomposition from image collections. In: International Conference on Computer Vision. pp. 12684–12694 (2021). https://doi.org/10.1109/ICCV48922.2021.01245
6. Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., Van Gool, L.: Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 745–755 (2022). https://doi.org/10.1109/CVPRW56347.2022.00090
7. Camuffo, E., Barbato, F., Ozay, M., Milani, S., Michieli, U.: Mocha: Multi-modal objects-aware cross-architecture alignment. In: European Conference on Computer Vision (2026)
8. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 19457–19467 (2024). https://doi.org/10.1109/CVPR52733.2024.01840
9. Chen, A., Xu, Z., Wei, X., Tang, S., Su, H., Geiger, A.: Dictionary fields: Learning a neural basis decomposition. ACM Transactions on Graphics (2023), https://doi.org/10.1145/3592135
10. Chen, Q., Shu, S., Bai, X.: Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis. In: European Conference on Computer Vision. pp. 253–269. Springer (2024). https://doi.org/10.1007/978-3-031-73383-3_15
11. Chen, Y., Wu, Q., Zheng, C., Cham, T.J., Cai, J.: Sem2nerf: Converting single-view semantic masks to neural radiance fields. In: European Conference on Computer Vision. pp. 730–748. Springer (2022). https://doi.org/10.1007/978-3-031-19781-9_42
12. Chou, Z.T., Huang, S.Y., Liu, I., Wang, Y.C.F., et al.: Gsnerf: Generalizable semantic neural radiance fields with enhanced 3d scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 20806–20815 (2024). https://doi.org/10.1109/CVPR52733.2024.01966
13. Dai, P., Xu, J., Xie, W., Liu, X., Wang, H., Xu, W.: High-quality surface reconstruction using gaussian surfels. In: ACM International Conference on Computer

Graphics and Interactive Techniques. pp. 1–11 (2024). https://doi.org/10.1145/3641519.3657441
14. Darling, B.A., Ferwerda, J.A., Berns, R.S., Chen, T.: Real-time multispectral rendering with complex illumination. In: Color Imaging Conference. vol. 19, pp. 345–351. Society of Imaging Science and Technology (2011). https://doi.org/10.1145/3721250.3743035
15. Dave, A., Zhao, Y., Veeraraghavan, A.: Pandora: Polarization-aided neural decomposition of radiance. In: European Conference on Computer Vision. pp. 538–556. Springer (2022). https://doi.org/10.1007/978-3-031-20071-7_32
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
17. Gibson, S.F.: Using distance maps for accurate surface representation in sampled volumes. In: IEEE Symposium on Volume Visualization and Graphics. pp. 23–30 (1998)
18. Griffiths, R., Dansereau, D.G.: RoRE: Rotary ray embedding for generalised multimodal scene understanding. In: International Conference on Learning Representations (2026), https://openreview.net/forum?id=BR2ItBcqOo
19. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: International Conference on Machine Learning (2020), https://dl.acm.org/doi/abs/10.5555/3524938.3525293
20. Großmann, W., Horn, H., Niggemann, O.: Improving remote material classification ability with thermal imagery. Scientific Reports **12**(1), 17288 (2022). https://doi.org/10.1038/s41598-022-21588-4
21. Guo, H., Liu, H., Wen, J., Li, J.: Cross-spectral gaussian splatting with spatial occupancy consistency. In: AAAI Conference on Artificial Intelligence. vol. 39, pp. 3229–3237 (2025). https://doi.org/10.1609/aaai.v39i3.32333
22. Han, Y., Tie, B., Guo, H., Lyu, Y., Li, S., Shi, B., Jia, Y., Ma, Z.: Polgs: Polarimetric gaussian splatting for fast reflective surface reconstruction. In: International Conference on Computer Vision. pp. 28073–28082 (2025)
23. Hashimoto, N., Murakami, Y., Bautista, P.A., Yamaguchi, M., Obi, T., Ohyama, N., Uto, K., Kosugi, Y.: Multispectral image enhancement for effective visualization. Optics Express **19**(10), 9315–9329 (2011). https://doi.org/10.1364/OE.19.009315
24. Hassan, M., Forest, F., Fink, O., Mielle, M.: Thermonerf: A multimodal neural radiance field for joint rgb-thermal novel view synthesis of building facades. Advanced Engineering Informatics **65**, 103345 (2025). https://doi.org/10.1016/j.aei.2025.103345
25. He, H., Liang, Y., Xiao, S., Chen, J., Chen, Y.: Cp-nerf: Conditionally parameterized neural radiance fields for cross-scene novel view synthesis. Computer Graphics Forum **42** (2023). https://doi.org/10.1111/cgf.14940
26. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM International Conference on Computer Graphics and Interactive Techniques. pp. 1–11 (2024). https://doi.org/10.1145/3641519.3657428
27. Huang, Y.H., He, Y., Yuan, Y.J., Lai, Y.K., Gao, L.: Stylizednerf: Consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 18321–18331 (2022). https://doi.org/10.1109/CVPR52688.2022.01780

28. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5885–5894 (2021). https://doi.org/10.1109/ICCV48922.2021.00583
29. Jin, H., Liu, I., Xu, P., Zhang, X., Han, S., Bi, S., Zhou, X., Xu, Z., Su, H.: Tensoir: Tensorial inverse rendering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2023). https://doi.org/10.1109/CVPR52729.2023.00024
30. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023). https://doi.org/10.1145/3592433
31. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: International Conference on Computer Vision. pp. 19729–19739 (2023). https://doi.org/10.1109/ICCV51070.2023.01807
32. Kim, Y., Jin, W., Cho, S., Baek, S.H.: Neural spectro-polarimetric fields. In: ACM International Conference on Computer Graphics and Interactive Techniques - Asia. pp. 1–11 (2023). https://doi.org/10.1145/3610548.3618172
33. Lei, C., Huang, X., Zhang, M., Yan, Q., Sun, W., Chen, Q.: Polarized reflection removal with perfect alignment in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1750–1758 (2020). https://doi.org/10.1109/CVPR42600.2020.00182
34. Lei, C., Qi, C., Xie, J., Fan, N., Koltun, V., Chen, Q.: Shape from polarization for complex scenes in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 12632–12641 (2022). https://doi.org/10.1109/CVPR52688.2022.01230
35. Li, C., Ono, T., Uemori, T., Mihara, H., Gatto, A., Nagahara, H., Moriuchi, Y.: Neisf: Neural incident stokes field for geometry and material estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 21434–21445 (2024). https://doi.org/10.1109/CVPR52733.2024.02025
36. Li, C., Ono, T., Uemori, T., Nitta, S., Mihara, H., Gatto, A., Nagahara, H., Moriuchi, Y.: Neisf++: Neural incident stokes field for polarized inverse rendering of conductors and dielectrics. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 26493–26503 (2025). https://doi.org/10.1109/CVPR52734.2025.02467
37. Li, J., Li, Y., Sun, C., Wang, C., Xiang, J.: Spec-nerf: Multi-spectral neural radiance fields. In: International Conference on Acoustics, Speech, and Signal Processing. pp. 2485–2489. IEEE (2024). https://doi.org/10.1109/ICASSP48485.2024.10446015
38. Li, R., Liu, J., Liu, G., Zhang, S., Zeng, B., Liu, S.: Spectralnerf: Physically based spectral rendering with neural radiance field. In: AAAI Conference on Artificial Intelligence. vol. 38, pp. 3154–3162 (2024). https://doi.org/10.1609/aaai.v38i4.28099
39. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (2023). https://doi.org/10.1109/CVPR52729.2023.00817
40. Liang, Z., Zhang, Q., Feng, Y., Shan, Y., Jia, K.: Gs-ir: 3d gaussian splatting for inverse rendering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 21644–21653 (2024). https://doi.org/10.1109/CVPR52733.2024.02045

41. Lincetto, F., Agresti, G., Rossi, M., Zanuttigh, P.: Exploiting multiple priors for neural 3d indoor reconstruction. In: British Machine Vision Conference. BMVA (2023)
42. Lincetto, F., Agresti, G., Rossi, M., Zanuttigh, P.: Multimodalstudio: A heterogeneous sensor dataset and framework for neural rendering across multiple imaging modalities. In: IEEE Conference on Computer Vision and Pattern Recognition (2025). https://doi.org/10.1109/CVPR52734.2025.01024
43. Liu, Y., Hu, B., Huang, J., Tai, Y.W., Tang, C.K.: Instance neural radiance field. In: International Conference on Computer Vision. pp. 787–796 (October 2023). https://doi.org/10.1109/ICCV51070.2023.00079
44. Lu, R., Chen, H., Zhu, Z., Qin, Y., Lu, M., Yan, C., et al.: Thermalgaussian: Thermal 3d gaussian splatting. In: International Conference on Learning Representations (2025)
45. Ma, R., Ma, T., Guo, D., He, S.: Novel view synthesis and dataset augmentation for hyperspectral data using nerf. IEEE Access **12**, 45331–45341 (2024). https://doi.org/10.1109/ACCESS.2024.3381531
46. Meng, G., Cai, Z., Chen, R., Tu, J., Wang, Y., Huang, Y., Ding, X.: Frn: Fractal-based recursive spectral reconstruction network. In: Advances in Neural Information Processing Systems (2025)
47. 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: European Conference on Computer Vision (2020), https://doi.org/10.1007/978-3-030-58452-8_24
48. Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4m: Massively multimodal masked modeling. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 58363–58408. Curran Associates, Inc. (2023), https://dl.acm.org/doi/10.5555/3666122.3668666
49. Osher, S., Fedkiw, R., Piechor, K.: Level set methods and dynamic implicit surfaces. Applied Mechanics Reviews **57**(3), B15–B15 (2004), https://doi.org/10.1007/b98879
50. Özer, M., Weiherer, M., Hundhausen, M., Egger, B.: Exploring multi-modal neural scene representations with applications on thermal imaging. In: European Conference on Computer Vision. pp. 82–98. Springer (2024). https://doi.org/10.1007/978-3-031-92805-5_6
51. Perez, F., Rojas, S., Hinojosa, C., Rueda-Chacón, H., Ghanem, B.: Unmix-nerf: Spectral unmixing meets neural radiance fields. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 26284–26293 (2025)
52. Poggi, M., Ramirez, P.Z., Tosi, F., Salti, S., Mattoccia, S., Di Stefano, L.: Cross-spectral neural radiance fields. In: International Conference on 3D Vision. pp. 606–616. IEEE (2022). https://doi.org/10.1109/3DV57658.2022.00071
53. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024). https://doi.org/10.1109/CVPR52733.2024.01895
54. Qu, Y., Dai, S., Li, X., Lin, J., Cao, L., Zhang, S., Ji, R.: Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In: ACM International Conference on Multimedia. pp. 5328–5337 (2024). https://doi.org/10.1145/3664647.3680852
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer

Vision and Pattern Recognition. pp. 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
56. Saponaro, P., Sorensen, S., Kolagunda, A., Kambhamettu, C.: Material classification with thermal imagery. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4649–4656 (2015). https://doi.org/10.1109/CVPR.2015.7299096
57. Shi, Z., Chen, C., Xiong, Z., Liu, D., Wu, F.: Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop. pp. 939–947 (2018). https://doi.org/10.1109/CVPRW.2018.00139
58. Thirgood, C., Mendez, O., Ling, E., Storey, J., Hadfield, S.: Hypergs: Hyperspectral 3d gaussian splatting. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5970–5979 (2025). https://doi.org/10.1109/CVPR52734.2025.00560
59. Tsagaris, V., Anastassopoulos, V.: Multispectral image fusion for improved rgb representation based on perceptual attributes. International Journal of Remote Sensing **26**(15), 3241–3254 (2005). https://doi.org/10.1080/01431160500127609
60. Varma, M., Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z.: Is attention all that nerf needs? In: International Conference on Learning Representations (2023)
61. Wang, H., Wen, S., Guo, B.: Polarimetric monocular gaussian splatting slam for dense surface reconstruction. In: ACM International Conference on Multimedia. pp. 7519–7528 (2025). https://doi.org/10.1145/3746027.3754925
62. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems (2021), https://dl.acm.org/doi/10.5555/3540261.3542342
63. Wang, W., Zhang, J., Shen, C.: Improved human detection and classification in thermal images. In: IEEE International Conference on Image Processing. pp. 2313–2316. IEEE (2010). https://doi.org/10.1109/ICIP.2010.5649946
64. Wang, Y., Huang, D., Ye, W., Zhang, G., Ouyang, W., He, T.: Neurodin: A two-stage framework for high-fidelity neural surface reconstruction. Advances in Neural Information Processing Systems **37**, 103168–103197 (2024), https://dl.acm.org/doi/10.5555/3737916.3741194
65. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: International Conference on Computer Vision. pp. 3295–3306 (2023). https://doi.org/10.1109/ICCV51070.2023.00305
66. Wang, Y., Fang, L., Zhu, H., Hu, F., Ye, L., Ma, Z.: Golf-nrt: Integrating global context and local geometry for few-shot view synthesis*. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 21349–21359 (2025). https://doi.org/10.1109/CVPR52734.2025.01989
67. Wu, S., Basu, S., Broedermann, T., Van Gool, L., Sakaridis, C.: Pbr-nerf: Inverse rendering with physics-based neural fields. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10974–10984 (2025). https://doi.org/10.1109/CVPR52734.2025.01025
68. Wu, Y., Dian, R., Li, S.: Multistage spatial-spectral fusion network for spectral super-resolution. IEEE Transactions on Neural Networks and Learning Systems **36**(7), 12736–12746 (2024). https://doi.org/10.1109/TNNLS.2024.3460190

69. Xu, J., Liao, M., Kathirvel, R.P., Patel, V.M.: Leveraging thermal modality to enhance reconstruction in low-light conditions. In: European Conference on Computer Vision. pp. 321–339. Springer (2024). https://doi.org/10.1007/978-3-031-72913-3_18
70. Xu, Y., Zoss, G., Chandran, P., Gross, M., Bradley, D., Gotardo, P.: Renerf: Relightable neural radiance fields with nearfield lighting. In: International Conference on Computer Vision. pp. 22581–22591 (2023). https://doi.org/10.1109/ICCV51070.2023.02064
71. Yao, M., Wang, M., Tam, K.M., Li, L., Xue, T., Gu, J.: Polarfree: Polarization-based reflection-free imaging. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10890–10899 (2025). https://doi.org/10.1109/CVPR52734.2025.01017
72. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Advances in Neural Information Processing Systems (2021), https://dl.acm.org/doi/10.5555/3540261.3540628
73. Ye, T., Wu, Q., Deng, J., Liu, G., Liu, L., Xia, S., Pang, L., Yu, W., Pei, L.: Thermal-nerf: Neural radiance fields from an infrared camera. In: International Conference on Intelligent Robots and Systems. pp. 1046–1053. IEEE (2024)
74. Yin, Q., Guo, P.: Multispectral remote sensing image classification with multiple features. In: International Conference on Machine Learning and Cybernetics. vol. 1, pp. 360–365. IEEE (2007). https://doi.org/10.1109/ICMLC.2007.4370170
75. Zhang, D., Yuan, Y.J., Chen, Z., Zhang, F.L., He, Z., Shan, S., Gao, L.: Stylizedgs: Controllable stylization for 3d gaussian splatting. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(12), 11961–11973 (2025). https://doi.org/10.1109/TPAMI.2025.3604010
76. Zhang, K., Luan, F., Li, Z., Snavely, N.: Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5565–5574 (2022). https://doi.org/10.1109/CVPR52688.2022.00548
77. Zhang, K., Lyu, Y., Guo, H., Li, S., Ma, Z., Shi, B.: Polaranything: Diffusion-based polarimetric image synthesis. In: International Conference on Computer Vision. pp. 26466–26476 (2025)
78. 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). https://doi.org/10.1109/ACCESS.2023.3254548
79. Zhang, Y., Chen, A., Wan, Y., Song, Z., Yu, J., Luo, Y., Yang, W.: Ref-gs: Directional factorization for 2d gaussian splatting. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 26483–26492 (2025). https://doi.org/10.1109/CVPR52734.2025.02466
80. Zhao, C., Huang, X., Yang, K., Wang, X., Wang, Q.: Generalizable 3d gaussian splatting for novel view synthesis. Pattern Recognition **161**, 111271 (2025)
81. Zhou, Z., Dong, M., Xie, X., Gao, Z.: Fusion of infrared and visible images for night-vision context enhancement. Applied Optics **55**(23), 6480–6490 (2016). https://doi.org/10.1364/AO.55.006480
82. Zhu, H., Ding, T., Chen, T., Zharkov, I., Nevatia, R., Liang, L.: Caesarnerf: Calibrated semantic representation for few-shot generalizable neural rendering. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) European Conference on Computer Vision. pp. 71–89. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72658-3_5