Beyond Appearances: Material Segmentation with Embedded Spectral Information from RGB-D imagery

Fabian Perez
Universidad Industrial de Santander
nelson2200183@correo.uis.edu.co

Hoover Rueda-Chacón
Universidad Industrial de Santander
hfarueda@uis.edu.co

Abstract

In the realm of computer vision, material segmentation of natural scenes represents a challenge, driven by the complex and diverse appearances of materials. Traditional approaches often rely on RGB images, which can be deceptive given the variability in appearances due to different lighting conditions. Other methods, that employ polarization or spectral imagery, offer a more reliable material differentiation but their cost and accessibility restrict their everyday usage. In this work, we propose a deep learning framework that bridges the gap between high-fidelity material segmentation and the practical constraints of data acquisition. Our approach leverages a training strategy that employs a paired RGBD-spectral data to incorporate spectral information directly within the neural network. This encoding process is facilitated by a Spectral Feature Mapper (SFM) layer, a novel module that embeds unique spectral characteristics into the network, thus enabling the network to infer materials from standard RGB-D images. Once trained, the model allows to conduct material segmentation on widely available devices without the need for direct spectral data input. In addition, we generate the 3D point cloud from the RGB-D image pair, to provide a richer spatial context for scene understanding. Through simulations using available datasets, and real experiments conducted with an iPad Pro, our method demonstrates superior performance in material segmentation compared to other methods. Code is available at: https://github.com/Factral/Spectral-material-segmentation

1. Introduction

Image segmentation consists on classifying pixels into multiple homogeneous regions, with each region exhibiting similar properties. In particular, material segmentation seeks to classify these pixels based on the objects' material rather than in terms of the objects themselves, as it is the case in semantic segmentation. This task is valuable as it provides the foundational blocks for various applications such as haptics [9], robotic navigation [31], acoustic simulation [2], and grasping [5]. Unlike semantic segmentation, material segmentation is more challenging since spectral reflectance signatures of objects are preferred over color information, for high reliability. However, acquiring spectral information is not an easy nor cheap task, and its mainstream usage is still restricted to laboratories or remote sensing platforms.

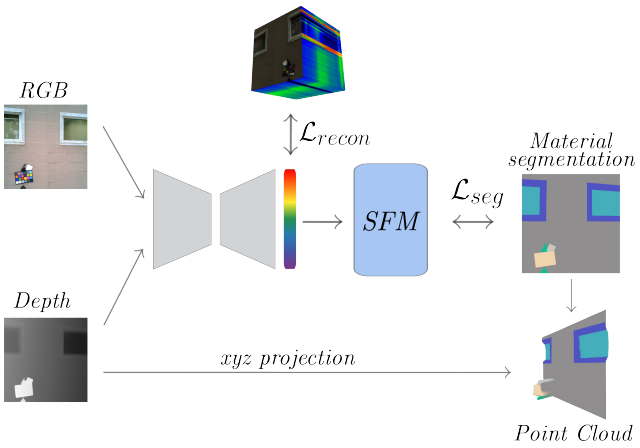

Figure 1. Overview of the training process and resulting output using an RGB-D input. The system employs $\mathcal{L}_{recon}$ for reconstruction of the spectral cube and $\mathcal{L}_{seg}$ for optimizing the segmentation task. A point cloud is then created using an xyz projection.

In contrast, RGB images are ubiquitous, and color sensors are within the reach of our hand. Nonetheless, extracting material from just 3 channels is challenging, if not impossible, and unreliable. With the advent of deep learning, some alternatives were proposed to do material segmentation from RGB images. Examples include UPerNet [33], which integrates object detection, semantic segmentation, and material segmentation to enhance overall performance, DMS-25 [28] that leverages the largest densely annotated material segmentation dataset to date, and MAC-

CNN [26], which focuses on local patch material recognition. These works have contributed valuable insights, collectively exhibiting moderate performance when compared to their semantic segmentation counterparts, a field nearing resolution through advances like SAM [13] and SegFormer [34]. This performance gap can be attributed to two primary challenges: the scarcity of large annotated datasets for material segmentation and the complexity of material appearance classification, exacerbated by phenomena such as metamerism, where identical RGB values may represent different materials. Addressing this gap, recent works have explored alternative data modalities such as, the use of infrared and polarization data in multimodal material segmentation [16], and hyperspectral imagery for enhanced road material segmentation [20], demonstrating significant improvements. However, these methods are often impractical for everyday use due to the prohibitive cost and complexity of the required equipment.

A recent work, coined MatSpecNet [12], has made strides by proposing a physically-constrained reconstruction of hyperspectral images for comparison against existing material databases, along with semantic segmentation information for the task of material segmentation. Despite its innovation, MatSpecNet's complexity and the heavyweight nature of its components render it unsuitable for deployment on devices with limited resouces, highlighting a pressing need for simpler, yet effective solutions.

Motivated by this challenge, we propose a novel framework that leverages convolutional neural networks (CNNs) to conduct material segmentation from RGB images, assisted by the depth map image encountered in RGB-D imagery, and introduces a new module, the Spectral Feature Mapper (SFM), designed to integrate spectral information into encoder-decoder architectures, as depicted in Figure 1, significantly enhancing material segmentation performance. By simplifying the integration of spectral embeddings into a model trainable with only RGB data, our approach relies on spectral data paired with each of the RGB images, but just for training. At evaluation time, only the RGB-D image pair is required. In particular, we use the Light Industrial Building HSI (LIB-HSI) dataset [11] for simulations and training. To test the capabilities of the proposed approach in real experiments, we use an iPad Pro and its built-in sensors to acquire a set of RGB-D image pairs of a scene of interest. Our approach promises to not only bridge the gap between high-fidelity material segmentation and practical application constraints but also looks forward to pave the way for widespread adoption of advanced material segmentation techniques across a variety of consumer electronics devices. This initiative marks a pivotal step towards the democratization of high-precision material segmentation, offering a robust solution to one of the field's most pressing problems.

2. Related Work

Material Segmentation. Inception work on material segmentation focused on recognition, which involved classifying entire images into one of several material categories. In recent years, the trend has shifted towards dense material segmentation, i.e pixel-wise classification on images. Typical strategies can be grouped into two main categories: those based on RGB data and the non-RGB-based modalities. RGB-based methods exploit annotated datasets such as OpenSurfaces [3] with 19,000 images across 37 materials, the Materials in Context Database [4] with 400,000 images with patch-level labels for 23 materials, and the Local Materials Database [26] with 5,000 images across 16 classes and include a tree hierarchy to structure material labels. A work used these datasets with a CNN for pixel-level classification complemented by a semantic segmentation network for additional global context clues [25]. A limitation of these approaches is their sparse labeling, with not all the pixels being densely annotated, due to the dataset used. To alleviate these problems, more recently, the Dense Material Segmentation (DMS) dataset [28] was proposed, featuring an extensive collection of 45,000 images across 52 densely labeled classes.

On the other hand, non-RGB methods have shown that additional information from different sensors can enhance the segmentation and make it more robust. For instance, the use of the Bidirectional Reflectance Distribution Function (BRDF) has been notable for its unique association of visual appearance with materials [30, 35]. While offering distinct advantages, measuring the BRDF demands a controlled environment, significant time, and high computational resources. Other approaches combines multiple modalities including RGB, infrared and polarization measurements [16]. This combination of data enriches the information available for segmentation, leading to improved results. Another modality uses spectral information for segmentation, exploiting the fact that each material exhibits a unique spectral signature, thus, enabling discrimination based on material properties. A notable study [38] developed its own hyperspectral material dataset combining five typical hyperspectral datasets, with manual material labeling added to them, resulting in a dataset spanning 28 spectral channels from 430 to 700nm at 10nm intervals. They employed support vector machines (SVM) for segmentation, and a CNN in a follow-up work [39]. It is important to note that, unfortunately, this unique dataset is no longer accessible, posing challenges for ongoing research and replication. In line with this spectral approach, [24] introduced the use of learned spectral filters for pixel-level material classification. This method is limited by environmental variations and depends on a complex optical setup.

RGB-D Image Acquisition. Efforts to acquire spectral information along with depth maps of the same scene

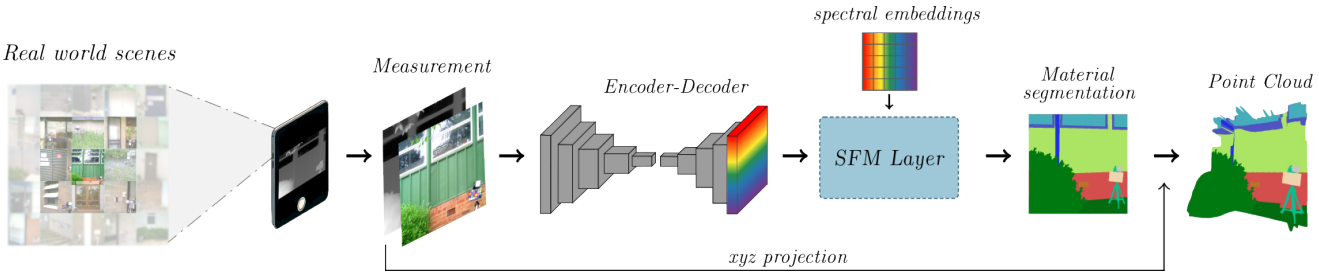

Figure 2. Illustration of the proposed framework utilizing the Spectral Feature Matching (SFM) layer for RGB-D image processing. An off-the-shelf device captures an RGB image along with depth information, which is then processed by an encoder-decoder network producing a hyperspectral reconstruction $\hat{\mathbf{H}}$. This reconstruction is operated on by the SFM layer using learned spectral embeddings to achieve material segmentation. Furthermore, a segmented point cloud is generated leveraging the captured depth data.

have given raise to RGB-D imaging systems, such as the Microsoft Kinect and the Asus Xtion sensor, which conventionally uses an RGB color sensor and a passive or active ranging sensor to estimate depth and spectral information, simultaneously. These systems are limited in their expressive spectral capabilities to only 3 channels, and the imaging modalities are carried out independently via multiple sensors. There exist two main modalities to acquire depth information, the one that relies on passive illumination, such as stereo [14] and light field imaging [32], and the one that requires active illumination, such as structured light imaging (SLI) [10], time-of-flight (ToF) imaging [15] and light detection and ranging (Lidar) [22] systems. Passive systems have the limitation that they cannot derive depth without depth cues created by texture. Active systems, in contrast, indirectly measure the round-trip propagation delay of a light pulse or the deformation of the pattern projected on the scene of interest to infer depth. Given the advantages of active illumination approaches, recent consumer electronic devices, such as smartphones and tablets, now incorporate this technology, thus making depth image acquisition mainstream. An example of a sensor equipped in these devices has shown a maximum sensing range of about 5 [27] and great robustness to lighting conditions [8], making it suitable for both indoor and outdoor sensing environments.

3. Proposed Method

Our approach for material segmentation introduces a novel framework tailored for deployment on off-the-shelf devices. It significantly enhances encoder-decoder architectures. The operational pipeline of our framework is depicted in Figure 2. In Section 3.1, we provide an overview of the framework, laying the foundation for understanding its operation. Section 3.2 delves into the specifics of the Spectral Feature Mapper (SFM) layer and its role in our model. Section 3.3 outlines the training loss formulation, critical for our method's success.

3.1. Framework Overview

The proposed deep learning framework integrates multiple data modalities to enhance material segmentation accuracy. It particularly uses spectral information to address misclassifications caused by the complex materials appearances [26]. The framework employs a dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{H}_i, \mathbf{D}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, with N samples, where $\mathbf{X}_i \in \mathbb{R}^{h\times w\times 3}$ represents the RGB image, $\mathbf{H}_i \in \mathbb{R}^{h\times w\times b}$ denotes the spectral image with b spectral bands, $\mathbf{D}_i \in \mathbb{R}^{h\times w}$ is the depth image, and $\mathbf{Y}_i \in \{1,\dots,c\}^{h\times w}$ represents the material segmentation labels for each pixel, with c being the number of classes; $h \times w$ represent the number of pixels along the vertical and horizontal extent, respectively.

In scenarios where RGB images are unavailable, their approximation can be achieved by leveraging the camera's spectral sensitivity $\phi_k(\lambda)$ for the RGB channels $k \in \{R, G, B\}$. This function interacts with the spectral reflectance of the scene, $r_{m,n}(\lambda)$, observed at each spatial location (m, n), $m = 0, 1, \dots, h$, $n = 0, 1, \dots, w$, where λ represents the wavelength. The approximation process calculates the intensity for each RGB channel and each intensity location, $I_{m,n,k}$, by

$$I_{m,n,k} = \int_{\lambda_{\min}}^{\lambda_{\max}} \phi_k(\lambda) \cdot r_{m,n}(\lambda)\, d\lambda. \tag{1}$$

In the discrete scenario $\phi_k(\lambda)$ can be modeled as $\mathbf{S} \in \mathbb{R}^{3\times b}$ and $r_{m,n}(\lambda)$ becomes $\mathbf{H}_i \in \mathbb{R}^{h\times w\times b}$. By reshaping the latter in matrix form, $\bar{\mathbf{H}} \in \mathbb{R}^{hw\times b}$, the operation can be expressed as

$$\bar{\mathbf{X}} = \bar{\mathbf{H}}\mathbf{S}^T \tag{2}$$

where $\bar{\mathbf{X}} \in \mathbb{R}^{hw\times 3}$, is then reshaped back to $\mathbf{X} \in \mathbb{R}^{h\times w\times 3}$ which represents the RGB image. This is done for all the N samples $\mathbf{H}_i$ in $\mathcal{D}$.

On the other hand, in scenarios where the depth image $\mathbf{D}_i$ is not present in $\mathcal{D}$, it can be generated through an inference process using a dedicated neural network tailored for

depth estimation from RGB [40], following

$$\mathbf{D}_i = f_{\Theta_d}(\mathbf{X}_i), \tag{3}$$

where f_{Θ_d} denotes the dedicated neural network function for depth estimation, parameterized by Θ_d, and $\mathbf{X}_i$ is the input RGB image.

The heart of our framework is anchored in an encoder-decoder network, f_Θ, designed to reconstruct hyperspectral images from RGB-D inputs; this is formalized as $f_\Theta : \mathbb{R}^{h\times w\times 4} \rightarrow \mathbb{R}^{h\times w\times b}$. These architectures have proven to be highly efficient for reconstruction [6, 37], owing to their ability to learn complex mappings from input to output spaces. The output of this reconstruction process is subsequently passed through the SFM layer, our proposed module, which improves the material segmentation prediction by leveraging embedded spectral information.

Once trained, our model only requires RGB-D images as input to predict the material segmentation map $\mathbf{y}_i$. From this segmentation and the depth map image $\mathbf{D}_i$, we can directly generate a point cloud $\mathcal{P} \in \{(x, y, z, \mathbf{y}_i)\}$, thereby translating depth into spatial coordinates (x, y, z) and associating each point with a material class $\mathbf{y}_i$. This capability inherently enhances scene understanding. A comprehensive illustration of our framework and its components is provided in Figure 2, which delineates the flow from input acquisition through material segmentation and to point cloud generation.

3.2. Spectral Feature Mapper (SFM) Layer

Inspired by the proven success of material classification strategies in remote sensing [19, 21], we introduce the SFM layer, a novel component seamlessly integrated into our deep learning framework. This layer is designed to universally enhance encoder-decoder architectures parameterized by f_Θ. The technique is grounded in the remote sensing practice of comparing an observed spectral signature against a database of known material spectral responses. The material that minimizes the spectral angle α to the observation is identified as the matching label. This approach, relying either on a comprehensive database of material responses or the use of linear unmixing, has demonstrated exceptional capabilities in distinguishing materials based on their unique spectral signatures [7, 23, 24].

The mathematical foundation of the SFM layer adapts the traditional approach to calculating spectral angles, treating the spectral signatures as vectors in a b-dimensional space; recall that b denotes the number of spectral bands. Unlike the standard method, our modification involves taking the negative of the spectral angle α between two matrices containing the spectral curves, $\hat{\mathbf{H}} \in \mathbb{R}^{h\times w\times b}$ (reconstructed spectrum) which is reshaped to $\bar{\hat{\mathbf{H}}} \in \mathbb{R}^{hw\times b}$ and

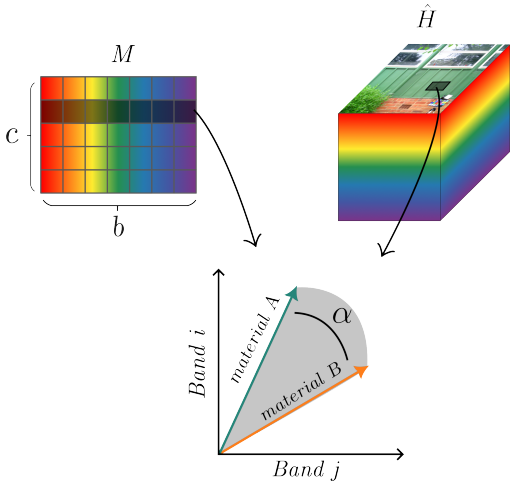

Figure 3. Visualization of the operation in the SFM layer, where $\hat{\mathbf{H}}$ denotes the reconstructed hyperspectral cube from f_θ, and $\mathbf{M}$ represents learned spectral embeddings. A pixel is selected to compare two spectral signatures; α denotes the similarity angle. This process is repeated for every pixel in the image, and the smallest α is selected indicating the material of the pixel.

$\mathbf{M} \in \mathbb{R}^{c\times b}$ (materials spectrum), as follows:

$$\alpha = -\cos^{-1}\left(\frac{\bar{\hat{\mathbf{H}}} \cdot \mathbf{M}^T}{\|\bar{\hat{\mathbf{H}}}\|\|\mathbf{M}\|}\right), \tag{4}$$

where $(\cdot)$ denotes the dot product, $\|\cdot\|$ is the Euclidean norm, $\alpha \in \mathbb{R}^{h\times w\times c}$ is the output angles. $\hat{\mathbf{H}}$ corresponds to the output of the encoder-decoder f_θ and $\mathbf{M}$ corresponds to the matrix of embeddings that will be learned and will contain the spectral information corresponding to each material. Figure 3 depicts the operation conducted in the SFM layer.

To ensure the correct operation of $\cos^{-1}$, it is essential to perform a clipping of its argument to the range $(-1, 1)$. This clipping step guarantees that the computed angles fall within the feasible domain of the inverse cosine, thereby avoiding undefined mathematical operations. Note that taking the negative of the cosine inverse transform, allows them to be treated as *logits* in the segmentation framework, where the softmax function and cross-entropy loss are typically employed, and calculated as,

$$\hat{\mathbf{y}} = \text{softmax}(\alpha). \tag{5}$$

Our approach guarantees that small angles, indicating closer spectral signature matches, are assigned higher probabilities after the softmax operation, in contrast to larger angles, which indicate a lesser degree of spectral signature similarity, and are thus attributed lower probabilities. This inversion of angle values allows for an intuitive integration into conventional segmentation workflows, where

Algorithm 1: PyTorch-style pseudocode for SFM layer

```
# h: output of encoder-decoder network
# m: embedding of materials

def forward(h):
    norm1 = m / sqrt(einsum('ij, ij → i', m, m)
    norm2 = sqrt(einsum('bijk, bijk → bij', h, h))
    dots = einsum('bij, mj → bim', h, norm1)
    dots = clamp(dots/norms2.unsqueeze(-1), -1, 1)
    # return the angles
    return -acos(dots).permute(0, 3, 1, 2)
```

higher *logits* lead to higher probabilities $\hat{\mathbf{y}}$. The detailed implementation and integration of this spectral angle calculation within our segmentation framework is further elaborated in Algorithm 1, where a PyTorch-style pseudocode is presented.

By incorporating this, we enable a direct comparison of spectral signatures of $\hat{\mathbf{H}}$ for material segmentation, effectively leveraging the unique spectral characteristics of materials as learnable parameters in $\mathbf{M}$ within the network. This dimensional structuring facilitates a comprehensive pixel-wise comparison across all pixels in the image, aligning with the network's segmentation loss. Furthermore, this approach is inherently compatible with backpropagation, as all operations involved are differentiable. Thus, the SFM layer can be seamlessly integrated into the training process, enabling the network to learn and update its parameters efficiently.

3.3. Training Schedule

The comprehensive training of our network involves the simultaneous training of the encoder-decoder architecture along with the SFM layer. To ensure training stability, a dummy ϵ is added to operations with a division, to avoid numerical division by zero errors. This modification is crucial for the robustness of the training process.

Additionally, the embedding matrix $\mathbf{M}$ is constrained within the interval $[0, 1]$. This constraint ensures compatibility with the reflectance values of $\mathbf{H}$, maintaining the θ values within the SFM layer align correctly for the accurate execution of spectral feature mapping operations. This is achieved through clipping at each training step, guaranteeing that the entire matrix remains within this specified range bounds. In particular, the clipping process is formalized as,

$$M_{i',j'} \leftarrow \min(\max(M_{i',j'}, 0), 1), \tag{6}$$

where i' indexes the material classes and j' indexes the spectral bands.

On the other hand, to optimize the performance of our network on the dual objective of material segmentation and hyperspectral image reconstruction, we define a composite loss function that incorporates both Mean Square Error (MSE) loss and Focal Loss (FL) [18]. The MSE component ensures that the output of the encoder-decoder network, f_Θ, closely matches the target hyperspectral images $\mathbf{H}$, thus guiding the spectral feature extraction towards high fidelity reconstruction. On the other hand, the FL component addresses the prevalent issue of class imbalance in material recognition datasets. By modifying the traditional cross-entropy loss to focus more intensively on hard-to-classify examples, the FL component ensures that our model is particularly adept at segmenting materials that are less frequently encountered and traditionally more challenging to classify.

The overall loss function of our network is a weighted sum of the FL for material segmentation $\mathcal{L}_{seg}$ and the MSE loss for hyperspectral image reconstruction $\mathcal{L}_{recon}$:

$$\mathcal{L} = \omega_1 \cdot \mathcal{L}_{seg}(\hat{\mathbf{y}}, \mathbf{y}) + \omega_2 \cdot \mathcal{L}_{recon}(\hat{\mathbf{H}}, \mathbf{H}), \tag{7}$$

$$\mathcal{L} = \omega_1 \cdot \text{FL}(\hat{\mathbf{y}}, \mathbf{y}) + \omega_2 \cdot \text{MSE}(\hat{\mathbf{H}}, \mathbf{H}), \tag{8}$$

where $\hat{\mathbf{y}}$ represents the predicted segmentation labels, $\mathbf{y}$ denotes the true labels, $\hat{\mathbf{H}}$ is the reconstructed hyperspectral output, $\mathbf{H}$ is the true hyperspectral cube, and ω_1, ω_2 are weights that balance the contribution of each loss component. Empirical evidence suggests that prioritizing segmentation loss ($\omega_1 > \omega_2$) enhances the model performance in material segmentation tasks. The training scheme is depicted in Figure 1.

This strategic approach to loss function design, allows us to simultaneously guide the network towards precise hyperspectral image reconstruction while effectively handling the complexities associated with material segmentation. The FL component, by focusing on hard-to-classify classes, ensures that all materials, regardless of their frequency in the dataset, are accurately segmented. In contrast, the MSE component ensures that the network's spectral reconstructions align closely with the true hyperspectral data, an important factor for the accurate interpretation of material properties.

4. Simulation Results

In this section, we evaluate the proposed framework on the publicly available Light Industrial Building HSI (LIB-HSI) dataset [11], providing a thorough assessment of our method. The scarcity of datasets that offer spectral-material segmentation, particularly in natural scenes, makes LIB-HSI a valuable resource for benchmarking state-of-the-art approaches in this domain.

Dataset. The LIB-HSI dataset comprises 44 categories of various materials across 513 facade images. While the dataset primarily focuses on building facades, the scenes are representative of natural environments. Accompanying each image, there is a hyperspectral datacube with 512x512 spatial pixels and 204 bands, ranging from 400 to 1000 nm. For our purpose, we kept the bands spanning 400 to 700

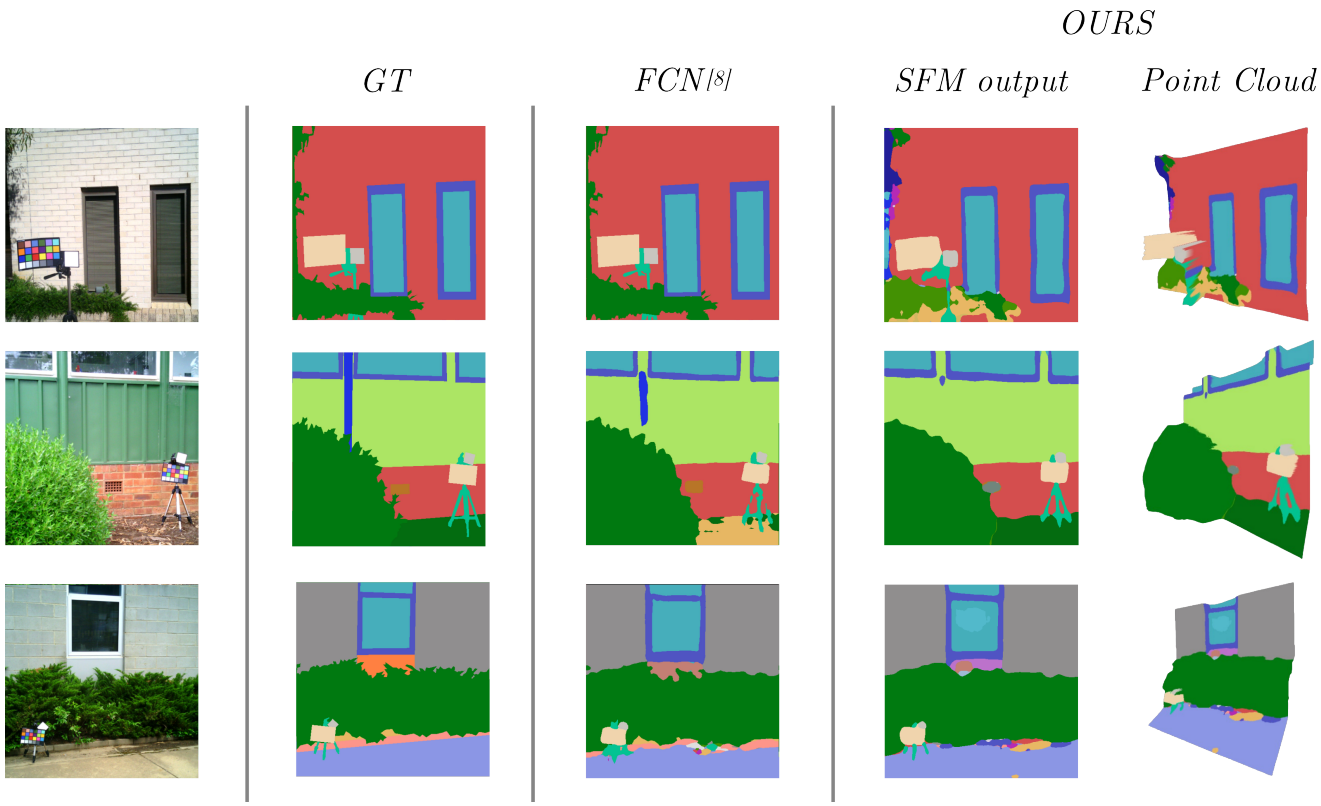

Figure 4. Visual comparison of material segmentation results. The first column displays the input RGB image, the second shows the ground truth for material segmentation, the third presents the results from the FCN method [11], and the fourth column illustrates our results with the predicted segmentation along with the segmented point cloud.

nm, and downsampled them to 31 bands. Further, each image is paired with a labeled materials map and a pseudo-RGB image. Given the absence of depth map information within the dataset, we simulated them using the depth-from-anything network [36] in its large version, which was employed solely for inference purposes on each dataset sample.

Architecture. For the encoder-decoder part of our network, we chose a pre-trained HRNet encoder [29] and a Feature Pyramid Network (FPN) decoder [17], known for their strong performance in object detection and segmentation tasks. At the output of this encoder-decoder network, the SFM layer was applied, and the entire network was trained.

Implementation Details. The framework was trained using the train set of LIB-HSI, using PyTorch on four NVIDIA T4 GPUs, with a batch size of 16, over 200 epochs. The weights for the loss function were set to $\omega_1 = 2$ and $\omega_2 = 0.5$, as these values demonstrated the best performance through cross-validation on the first 20 epochs. Data augmentation strategies included horizontal flipping, scaling, and rotation were used. A ReduceLROnPlateau strategy was employed to adjust the learning rate.

Evaluation and Metrics. To benchmark our method, we adhered to the same metrics employed in [28] and [11], that include pixel accuracy and mean intersection over union (IoU) over classes.

Input	Method	Accuracy	Average Class IoU
RGB	U-Net [11]	0.687	0.236
RGB	FCN [11]	0.829	0.443
RGB-D	Ours	**0.8647**	**0.4837**

Table 1. Comparison of results for different methods, in terms of classification accuracy (%) and mean IoU over classes, on the LIB-HSI test set.

We benchmark our proposed method against state-of-the-art approaches that have reported results on the LIB-HSI dataset. As demonstrated in Table 1, our method surpasses the existing baselines across both metrics, achieving a notable 0.8647 in accuracy and 0.4837 in mean IoU (mIoU). This performance leap underscores the advantage of incorporating spectral information into a deep learning framework for material segmentation in natural scenes. Previous

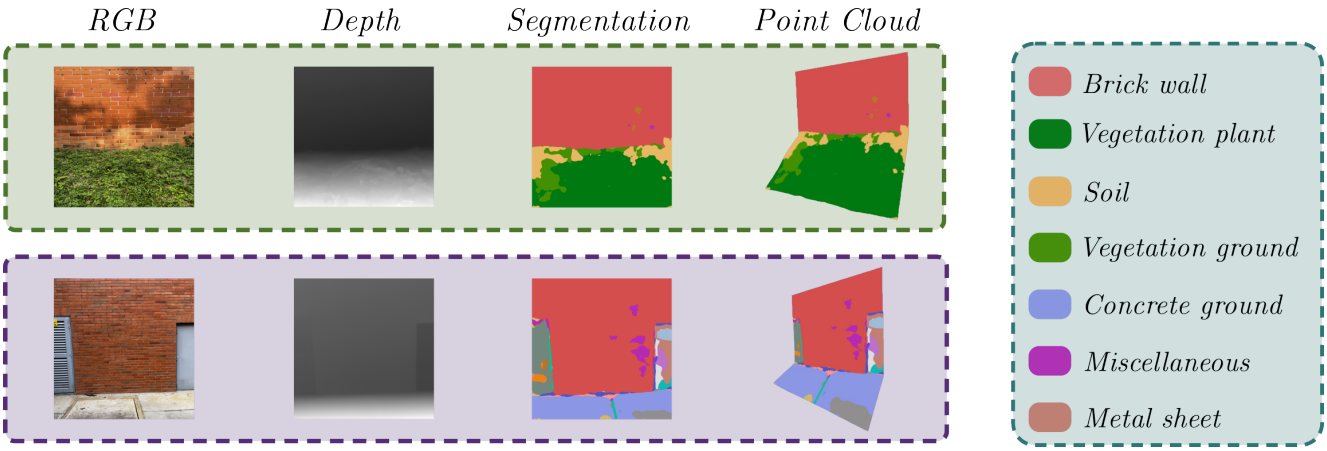

Figure 5. Visual outcomes from our experimental acquisition using an iPad Pro, displaying our framework's predictions trained on the LIB-HSI dataset. The point clouds are generated using the depth data extracted from the LiDAR.

approaches did not utilize the available spectral data $\mathbf{H}$, discarding valuable information and consequently leading to lower performance. In contrast, our framework leverages this spectral information through the SFM layer, thereby enhancing performance while retaining ease of capture on standard off-the-shelf devices.

In Figure 4, a visual comparison of our results is presented, illustrating that our method achieves great performance despite the deceptive appearances of some materials, outperforming previous techniques. Furthermore, for each prediction, we generate a point cloud utilizing the captured depth information, which significantly enhances the scene's comprehension.

One critical aspect of our method is that the matrix $\mathbf{M}$ is learned during training. Initially randomized, $\mathbf{M}$ converges to represent the spectral signatures of the materials over the course of training. This learning process allows $\mathbf{M}$ to embody a compact and discriminative spectral representation for each material class, enhancing the model's segmentation performance.

5. Experimental Results

To demonstrate the practicality and user-friendliness of our framework for everyday applications, we conducted experiments in a pair of real-world scenes resonant with the domain dataset. For this purpose, we employed an iPad Pro, leveraging its ability to capture high-quality RGB images and depth information through its LiDAR scanner using the ARKit API [1]. The official ARKit SDK was utilized to gather this data, providing depth information of up to 5 meters [27]. Figure 5 illustrates the results of these acquisitions, vividly demonstrating the efficacy of our proposed framework in accurately segmenting materials within diverse environments. Our model was capable of effectively identifying materials such as brick, vegetation, concrete, and metal under various lighting conditions, showcasing its robustness and adaptability. Through this empirical evaluation, we evaluated the significance and ease with which our framework can be applied to common environmental contexts.

6. Conclusions

This work presented a deep learning framework for material segmentation that integrates embedded spectral information within a standard RGB-D imaging workflow. Our method capitalizes on the encoded spectral characteristics through the Spectral Feature Mapper (SFM) layer, introducing a novel dimension to the segmentation task that extends beyond mere visual appearances. This is particularly advantageous in complex natural scenes where lighting conditions and surface textures may lead to ambiguous visual cues. Our approach demonstrates a significant advancement over traditional baseline RGB-based methods, effectively utilizing spectral information without the need for specialized sensor equipment. The versatility of our framework is underscored by its compatibility with readily available consumer electronic devices, like the iPad Pro, used in our real-world experiments. This accessibility paves the way for widespread adoption and practical deployments in diverse applications. The encouraging results make our framework a step forward in the realm of computer vision, bridging the gap between high-fidelity material recognition and practical application.

Acknowledgments

We thank the Vicerrectoría de Investigación y Extensión of the Universidad Industrial de Santander (VIE-UIS) for supporting this work registered with VIE code 3759.

References

[1] https://developer.apple.com/documentation/arkit. 7

[2] Anurag Arnab, Michael Sapienza, Stuart Golodetz, Julien Valentin, Ondrej Miksik, Shahram Izadi, and Philip Torr. Joint object-material category segmentation from audio-visual cues. *arXiv preprint arXiv:1601.02220*, 2016. 1

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32(4), 2013. 2

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 2

[5] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics*, 30(2):289–309, 2013. 1

[6] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 745–755, 2022. 4

[7] Sujata Chakravarty, Bijay Kumar Paikaray, Rutuparnna Mishra, and Satyabrata Dash. Hyperspectral image classification using spectral angle mapper. In *2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 87–90. IEEE, 2021. 4

[8] Lucía Díaz Vilariño, H Tran, Ernesto Frías Nores, Jesús Balado Frías, Kourosh Khoshelham, et al. 3d mapping of indoor and outdoor environments using apple smart devices. *ISPRS-International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, 2022. 3

[9] Yang Gao, Lisa Anne Hendricks, Katherine J Kuchenbecker, and Trevor Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 536–543. IEEE, 2016. 1

[10] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Advances in optics and photonics*, 3(2):128–160, 2011. 3

[11] Nariman Habili, Ernest Kwan, Weihao Li, Christfried Webers, Jeremy Oorloff, Mohammad Ali Armin, and Lars Petersson. A hyperspectral and rgb dataset for building façade segmentation. In *European Conference on Computer Vision*, pages 258–267. Springer, 2022. 2, 5, 6

[12] Yuwen Heng, Yihong Wu, Jiawen Chen, Srinandan Dasmahapatra, and Hansung Kim. Matspectnet: Material segmentation network with domain-aware and physically-constrained hyperspectral reconstruction. *arXiv preprint arXiv:2307.11466*, 2023. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[14] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2189–2199, 2020. 3

[15] Larry Li et al. Time-of-flight camera—an introduction. *Technical white paper*, (SLOA190B), 2014. 3

[16] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022. 2

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[19] Xiaofang Liu and Chun Yang. A kernel spectral angle mapper algorithm for remote sensing image classification. In *2013 6th International Congress on Image and Signal Processing (CISP)*, pages 814–818, 2013. 4

[20] Dhanushka C Liyanage, Robert Hudjakov, and Mart Tamre. Hyperspectral imaging methods improve rgb image semantic segmentation of unstructured terrains. In *2020 International Conference Mechatronic Systems and Materials (MSM)*, pages 1–5. IEEE, 2020. 2

[21] Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42(8):1778–1790, 2004. 4

[22] Thinal Raj, Fazida Hanim Hashim, Aqilah Baseri Huddin, Mohd Faisal Ibrahim, and Aini Hussain. A survey on lidar scanning mechanisms. *Electronics*, 9(5):741, 2020. 3

[23] Venkat Ravikiran S Rashmi S, Swapna Addamani. Spectral angle mapper algorithm for remote sensing image classification. In *2014 International Journal of Innovative Science, Engineering Technology*. IJISET, 2014. 4

[24] Vishwanath Saragadam and Aswin C Sankaranarayanan. Programmable spectrometry: Per-pixel material classification using learned spectral filters. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2020. 2, 4

[25] Gabriel Schwartz and Ko Nishino. Material recognition from local appearance in global context. *arXiv preprint arXiv:1611.09394*, 2016. 2

[26] Gabriel Schwartz and Ko Nishino. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1981–1995, 2019. 2, 3

[27] Alessandra Spreafico, Filiberto Chiabrando, L Teppati Losè, and F Giulio Tonolo. The ipad pro built-in lidar sensor: 3d rapid mapping tests and quality assessment. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:63–69, 2021. 3, 7

[28] Paul Upchurch and Ransen Niu. A dense material segmentation dataset for indoor and outdoor scene parsing. In

European Conference on Computer Vision, pages 450–466. Springer, 2022. 1, 2, 6

[29] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 6

[30] Oliver Wang, Prabath Gunawardane, Steve Scher, and James Davis. Material classification using brdf slices. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2805–2811. IEEE, 2009. 2

[31] Justin Wilson, Nicholas Rewkowski, and Ming C Lin. Audio-visual depth and material estimation for robot navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9239–9246. IEEE, 2022. 1

[32] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017. 3

[33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 1

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 2

[35] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential angular imaging for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–773, 2017. 2

[36] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 6

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 4

[38] Yu Zhang, Cong Phuoc Huynh, Nariman Habili, and King Ngi Ngan. Material segmentation in hyperspectral images with minimal region perimeters. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 834–838. IEEE, 2016. 2

[39] Yu Zhang, King Ngi Ngan, Cong Phuoc Huynh, and Nariman Habili. Learning deep spatial-spectral features for material segmentation in hyperspectral images. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2017. 2

[40] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020. 4