

RESEARCH ARTICLE

Physics-informed multimodal learning for snapshot dental spectral reflectance prediction

Yujun Feng¹ | Yushan Guo¹ | Qingyi Liu¹ | Fei Liu² | Jiefei Shen² | Junfei Shen¹

¹College of Electronics and Information Engineering, Sichuan University, Chengdu, China

²Department of Prosthodontics, West China Hospital of Stomatology Sichuan University, Chengdu, China

Correspondence

Junfei Shen.

Email: shenjunfei@scu.edu.cn

Funding information

National Key Research and Development Program of China, Grant/Award Number: 2022YFC2410102; Sichuan Science and Technology Program, Grant/Award Number: 2025YFHZ0333; Research and Development Program, West China Hospital, Sichuan University, Grant/Award Number: RD-03-202408; Young Elite Scientists Sponsorship Program by China Association for Science and Technology, Grant/Award Number: 2022QNRC001; Sichuan "Tianfu Emei Plan" talent project, Grant/Award Number: A0103602

Abstract

Accurate color matching is essential to achieving aesthetically realistic outcomes in dental crown and bridge restorations. Traditional visual methods, however, are often affected by lighting variations and observer subjectivity. These limitations can lead to metamerism and inconsistent clinical outcomes. In this paper, we propose a deep learning-based snapshot spectral reflectance prediction framework by incorporating multimodal optical priors into a physics-informed modular reconstruction pipeline. Architecturally, it employs dual subnetworks integrated with attention mechanisms to enhance both spectral and spatial fidelity. A dedicated dataset consisting of over 4000 RGB–hyperspectral image pairs of dental samples is constructed under diverse lighting conditions using a custom-built acquisition setup, enabling robust reflectance restoration. Quantitative evaluations demonstrate that the mean squared error of the reconstructed spectral reflectance is 0.002, with the structural similarity index measure achieving 0.8496. The proposed spectral reflectance reconstruction framework is fast, reliable, and robust, offering a physically consistent and clinically applicable solution for accurate color matching. Furthermore, its modular design provides flexibility, suggesting potential for adaptation and broader applicability across various biomedical imaging domains.

KEYWORDS

color matching, hyperspectral reconstruction, physical priors

1 | INTRODUCTION

Spectral reflectance, as a fundamental property of tooth, arises from differences in atomic and molecular structures [1]. The differences in spectral reflectance correspond to variations in tooth composition. These variations ultimately result in changes in tooth color. Spectral imaging technology leverages this property and has been widely applied in clinical research. Applications include tumor boundary detection, neural

localization, and assessing dental structure health [2]. In tooth restoration, metamerism can cause problems. This phenomenon occurs when different spectra appear color-matched under a specific illumination, potentially leading to mismatches between natural teeth and restorative materials under changing lighting conditions. Accurately matching the color and texture of a patient's natural teeth with restorative teeth is challenging. Traditional color matching methods often struggle to achieve this. Spectral reflectance, however,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Quantitative Biology* published by John Wiley & Sons Australia, Ltd on behalf of Higher Education Press.

can objectively differentiate materials. By analyzing it, clinicians can achieve more precise and natural restorations. This approach leverages the rich physiological and pathological information in the teeth [3, 4]. In dental treatment, spectral imaging helps dentists diagnose diseases more accurately and promptly. It also supports more refined, personalized, and comprehensive clinical solutions.

Existing spectral imaging techniques can be classified into spatial scanning imaging, filter-based imaging, and snapshot imaging. The method of spatial scanning imaging captures images or spectral data from different positions of the target using physical movement or an optical scanning mechanism. The complete hyperspectral information is then synthesized from these measurements. The scanning methods include pushbroom scanning and whiskbroom scanning. Pushbroom scanning acquires images by moving a linear detector array. This allows the spectrometer to capture spectral information over a large area quickly. It is particularly suitable for remote sensing on spaceborne platforms [5–7]. Whiskbroom scanning uses a static detector to continuously collect data using a moving platform. It is ideal for high-resolution monitoring of relatively static scenes. However, it is not suitable for dynamic targets [8, 9]. Overall, spatial scanning is time-consuming and constrained by environmental conditions. It requires precise relative motion between the target and the imaging system. These requirements increase hardware complexity and limit practical applications [10].

Filter-based imaging relies on narrow-band or wide-band filters to modulate spectral light signals. The filtered light is then used to reconstruct hyperspectral images. Narrow-band filters are highly selective, allowing only light within a specific wavelength range to pass. They provide detailed spectral information and are commonly used in applications requiring precise spectral analysis, such as agricultural monitoring and environmental surveillance [11, 12]. Wide-band filters allow light within a specific wavelength range to pass through, helping to capture a broader spectrum of information. They are often suitable for applications that require capturing a large range of spectra, such as geological exploration and remote sensing monitoring [13, 14]. However, because these techniques typically rely on tunable filters, each filter transition requires a certain amount of time, thereby limiting its performance due to the scanning speed constraint.

Snapshot imaging allows the simultaneous acquisition of spectral information at multiple wavelengths of an object in a single exposure. This technology typically employs various encoding techniques, such as amplitude encoding, joint phase encoding, and wavelength encoding, to map light at different wavelengths into a single image. The encoded image is then decoded using algorithms to reconstruct the spectral image

[8, 15]. Coded aperture snapshot spectral imaging (CASSI) provides an efficient mechanism for recovering 3D spectral data from a single 2D measurement. The principle of CASSI is to capture a compressed measurement in a single snapshot. A reconstruction algorithm is then used to recover the 3D hyperspectral cube [16, 17]. The application of deep learning replaces the iterative algorithm with the deep neural network, which improves the efficiency of spectral reconstruction [18–20]. Yu et al. proposed a deep learning enabled compact CASSI system (R-CASSI) to obtain a compact CASSI system with high spatial performance [21]. CASSI enables high-quality hyperspectral capture but is hindered by bulky optical designs. Recent research studies also focus on hardware simplification, such as compact optical designs, programmable masks, and multi-functional single optical elements, which reduce system bulk while maintaining spectral fidelity [22–24]. However, these data-driven approaches tend to ignore the physics behind image formation, limiting their generalizability and interpretability.

To overcome above challenges, this paper proposes a physics-informed deep learning-based hyperspectral reflectance prediction framework from existing linear RGB images. The framework utilizes an RGB camera to capture raw paired images of both a white board and restored teeth under controlled illumination. After capturing the images, the raw sensor data undergoes image signal processing (ISP) to generate linear RGB images. The processed image pairs are then fed into a dual-subnetwork architecture. The first subnetwork estimates the spectral power distribution of the scene illumination from the linear RGB image of a white board with the known camera sensitivity. Its output, together with camera sensitivity as physical priors, is subsequently utilized by the second subnetwork to reconstruct the hyperspectral reflectance of the target tooth from the linear RGB image of tooth. The framework is designed for end-to-end inference, requiring only linear RGB inputs and outputting the reconstructed hyperspectral reflectance, making it highly accessible for real-world dental applications.

The main contributions of our paper are: (1) a dental color matching pipeline, leveraging deep learning neural networks to analyze linear RGB images, is proposed, which enables snapshot-based spectral reflectance prediction; (2) a task-specific dataset, involving more than 4000 RGB-hyperspectral image pairs, is constructed to support model training and evaluation, which enhances the spectral reconstruction robustness and reliability; and (3) based on physical principles, the dual subnetworks, which integrate multi-modal learning and attention mechanisms, are established for inverse imaging to reconstruct the dental spectral reflectance and thereby improve spectral reconstruction accuracy.

2 | RESULTS

2.1 | Experimental scheme

Figure 1 outlines the complete workflow of our proposed hyperspectral imaging and reconstruction method, encompassing data acquisition, computational reconstruction, and result evaluation.

As shown in Figure 1A, the data acquisition setup employs a beam-splitter-based imaging system to simultaneously capture an RGB image and a ground-truth hyperspectral image of a sample (tooth or calibration target) under controlled illumination. This co-registered setup enables precise calibration. A standard ColorChecker is also captured to facilitate camera characterization.

The core reconstruction pipeline is depicted in Figure 1B. It involves three sequentially connected neural network modules that incorporate physical priors: First, the camera spectral sensitivity estimation network (MSNet) estimates the spectral sensitivity of the RGB camera using the linear RGB image of the ColorChecker. Second, the illumination spectrum estimation network (ISNet) predicts the spectral power distribution of the light source. It takes the linear RGB image of a white reference board and the estimated camera sensitivity as input. Finally, the hyperspectral reflectance reconstruction network (HRNet) recovers the target's spectral

reflectance. It uses the sample's linear RGB image, guided by the estimated illumination spectrum and camera sensitivity from the previous modules. For brevity, we refer to these pre-trained networks as MSNet, ISNet, and HRNet throughout the paper.

Figure 1C illustrates the evaluation stage. The reconstructed hyperspectral reflectance is compared against the ground truth obtained from the spectral camera. Both quantitative metrics (e.g., mean squared error [MSE], structural similarity index measure [SSIM]) and qualitative visual comparisons are employed to analyze the reconstruction accuracy and spectral fidelity.

2.2 | Training results

During the neural network model training process, the changes in the loss value during iterations are recorded as an indication of the model's increasing accuracy. Additionally, the final loss value and predicted results on the test set are also recorded. Figure 2 shows the variation of loss values during the training process. From the total loss, it can be observed that the loss quickly converges and stabilizes after around 75 epochs. The loss between the predicted spectral reflectance reconstruction and the ground truth values is 0.15.

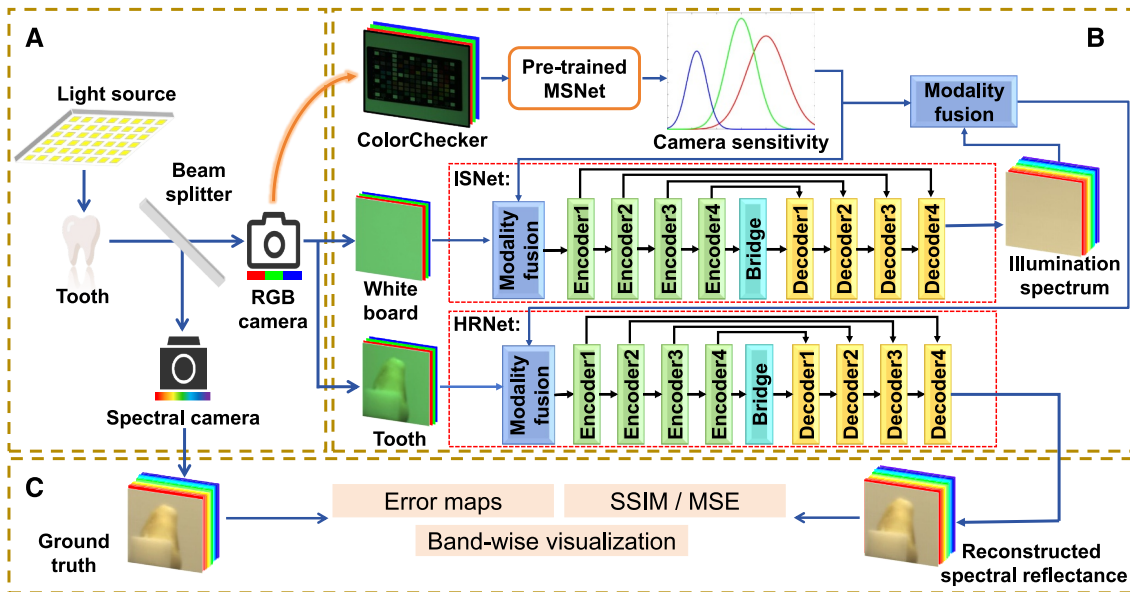


FIGURE 1 The overall experimental scheme of this paper. (A) The optical setup, incorporating a beam splitter, allows both RGB and spectral cameras to image the tooth or white board under the same illumination condition, facilitating subsequent registration and calibration. The ColorChecker is captured by an RGB camera. (B) The neural network framework comprises three subnetworks: a camera spectral sensitivity estimation subnetwork, which takes a ColorChecker's linear RGB image as input and estimates the camera spectral sensitivity; an illumination spectrum estimation subnetwork, which predicts the illumination spectrum from the linear RGB image of the white board, with the estimated camera spectral sensitivity; and a hyperspectral reflectance reconstruction subnetwork, which reconstructs the hyperspectral reflectance of the tooth from its linear RGB image, guided by the estimated camera spectral sensitivity and the illumination spectrum. (C) Comparison and analysis between reconstructed spectral reflectance and ground truth using MSE, SSIM, and visual inspection. MSE, mean squared error; SSIM, structural similarity index measure.

2.3 | Ablation study

To evaluate the contribution of attention mechanisms to spectral reconstruction, we conduct ablation experiments on both the illumination estimation subnetwork (ISNet) and the reflectance reconstruction subnetwork (HRNet). Specifically, we remove the channel attention (CA) and spatial attention (SA) modules from each subnetwork independently, and compare the results with the full model. These baseline variants are not intended to represent state-of-the-art methods, but are designed to isolate and quantify the impact of attention modules on reconstruction accuracy and perceptual fidelity.

Table 1 reports the MSE, CosLoss, and SSIM values for different network configurations. The proposed model (“Ours”), which integrates attention mechanisms in both subnetworks, achieves the overall performance with the lowest MSE (0.002) and the highest similarity index (0.8496). In contrast, removing all attention modules (“No Attn”) leads to significantly degraded performance, especially in SSIM, which drops to 0.8253. When attention is applied to only one subnetwork—either the ISNet or the HRNet—the performance improves compared to the no-attention baseline. These results demonstrate the necessity of incorporating attention in both branches to enhance spectral reconstruction accuracy.

To further validate the perceptual fidelity of the reconstructed spectra, we employed the CIEDE2000

color difference metric, which is widely used in dentistry and color science to quantify visual color differences. As shown in Figure 3, our method achieves the lowest mean ΔE value (0.95), which falls below the commonly accepted perceptibility threshold ($\Delta E < 1$) and indicates that the reconstructed colors are nearly indistinguishable from the ground truth to the human eye. In contrast, the baseline methods yield larger mean differences ($\Delta E = 1.38\text{--}1.51$), especially at the edges and high-texture regions of the tooth, where rapid spectral variations occur. This demonstrates that our approach provides superior perceptual color fidelity and enhances its potential clinical utility. The improvement can be attributed to the attention-based architecture, which enables the network to allocate more representational capacity to visually sensitive regions such as boundaries and textured surfaces. These areas exhibit steep spectral gradients, and conventional convolution-based baselines tend to smooth out such variations, resulting in larger perceptual discrepancies. By contrast, our method selectively preserves fine-grained spectral-spatial correlations, thereby reducing color mismatches in regions most critical for clinical shade assessment.

Figure 4 presents a comprehensive analysis of the prediction error distributions across different network configurations. The histogram (Figure 4A) illustrates the distribution of pixel-level reconstruction errors across the test set. The proposed model demonstrates the lowest overall error values, with the distribution skewed further left and concentrated below 0.04, indicating a higher reconstruction accuracy. The spatial error maps (Figure 4B) visualize the localized reconstruction errors. The model without attention exhibits noticeably higher error intensities along the boundaries between teeth and the standard whiteboard background compared to the full model (“Ours”), reflecting poorer reconstruction of object edges. These boundary-related errors are closely associated with regions of abrupt spectral-spatial transitions at object-background interfaces, where the reconstruction task involves a trade-off between preserving spectral fidelity and maintaining spatial continuity. Pixels near these boundaries have limited spatial context, making feature representation less reliable. The proposed model mitigates such artifacts by adaptively focusing on spectrally and spatially informative regions, though some residual error remains at extreme transitions.

Figure 5 presents grayscale visualizations of three representative spectral bands (570 nm, 530 nm, and 450 nm) reconstructed by different models. The proposed model exhibits the highest visual fidelity to the ground truth. The figure also reveals that all reconstruction results, except those from the proposed method, exhibit varying degrees of edge darkening. Because this artifact does not appear in the ground truth, it is likely introduced during the reconstruction

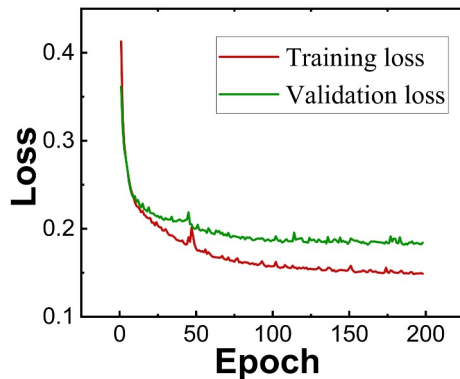


FIGURE 2 Training and validation loss curves over epochs.

TABLE 1 Specific values of MSE, cosine similarity loss (CosLoss) and SSIM values in different network structures.

Method	MSE	CosLoss	SSIM
No Attn	0.00343	0.00217	0.8253
HRNet without Attn	0.00230	0.00103	0.8450
ISNet without Attn	0.00225	0.00096	0.8469
Ours	0.00200	0.00090	0.8496

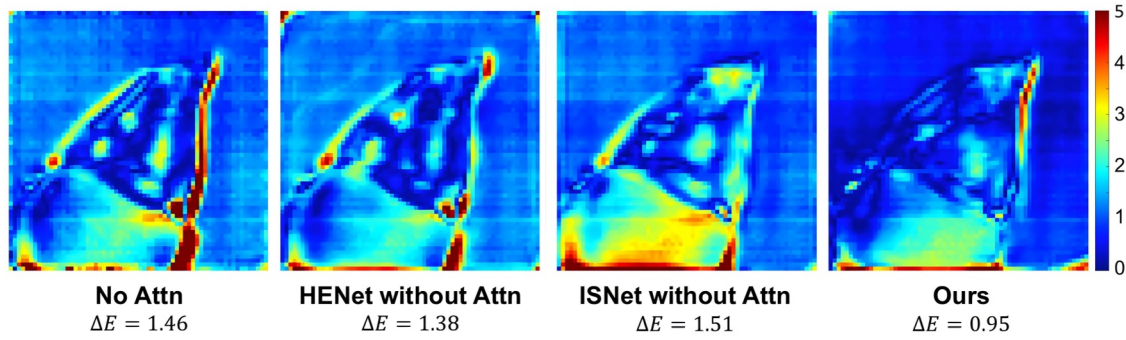


FIGURE 3 Visualization of the CIEDE2000 color difference maps between the ground truth and reconstructed spectra under different methods. Lower values indicate better perceptual color fidelity. Compared with other approaches ($\Delta E = 1.46$, 1.38 , and 1.51), our method achieves the lowest mean color difference ($\Delta E = 0.95$), suggesting that the reconstructed spectra are nearly indistinguishable from the ground truth in terms of human visual perception.

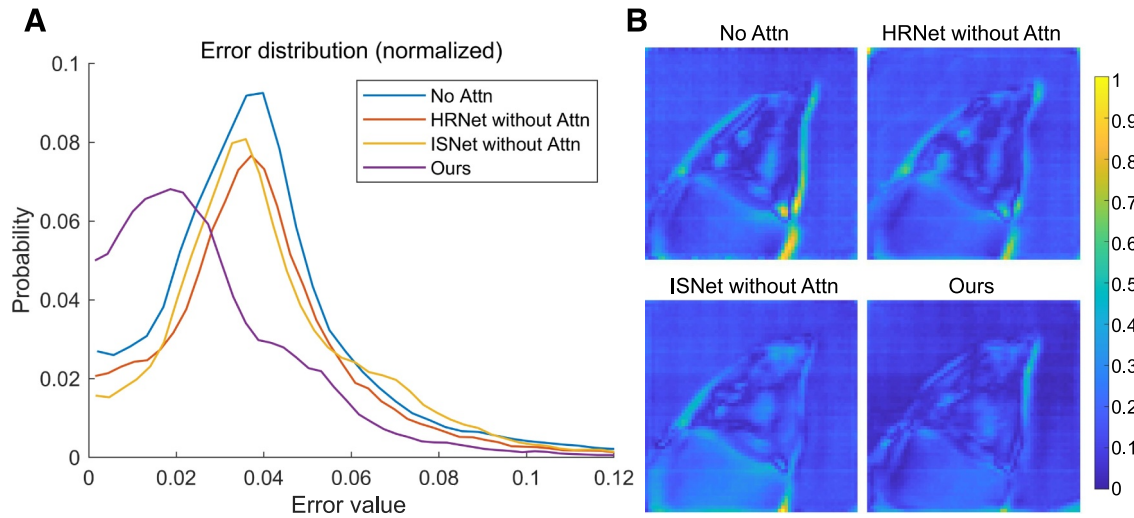


FIGURE 4 Both the statistical distribution and the spatial visualization of reconstruction errors are presented. (A) A line plot of the error distribution between the predicted and actual values for different network architectures. (B) Error images for different network structures.

process. A plausible explanation is that pixels near image boundaries have limited spatial context, making their spectral reconstruction more vulnerable to feature degradation. Networks without attention mechanisms are particularly affected, as they tend to allocate fewer representational resources to peripheral regions. These results suggest that attention mechanisms play a critical role in enhancing reconstruction at image boundaries by guiding the network to focus on informative regions, thereby mitigating boundary artifacts caused by insufficient feature representation at the periphery.

The MSE and SSIM results of the entire 32 spectral bands are illustrated in Figure 6. The proposed model consistently outperforms all baselines, achieving the lowest MSE and highest SSIM across nearly the entire spectral range. In particular, the benefit of attention mechanisms is most prominent in the 450–600 nm region with the MSE minimized.

In summary, the attention mechanisms in both ISNet and HRNet contribute complementary benefits. Their joint integration enables the model to focus more effectively on informative spectral–spatial cues, resulting in superior reconstruction performance.

3 | DISCUSSION

This paper presents a physics-informed multimodal hyperspectral reconstruction framework based on deep learning for dental color matching. By inverting the optical imaging process through a dual-network architecture, the proposed model enables accurate and efficient prediction of hyperspectral reflectance from linear RGB images acquired using snapshot-based scanning. In contrast to conventional color matching methods, this approach markedly alleviates dependence on expensive

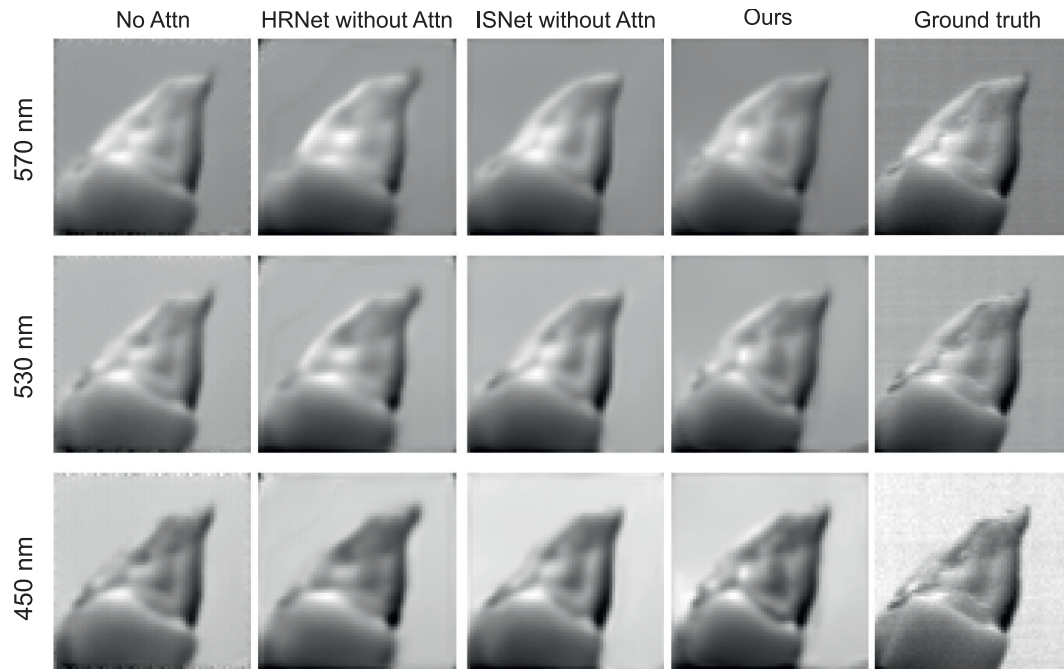


FIGURE 5 Comparison of reconstructions and true values for each different network architecture. Reconstructions are derived from single-channel grayscale images of the three selected bands (570 nm, 530 nm, 450 nm). All grayscale images are normalized to [0, 1] for visualization.

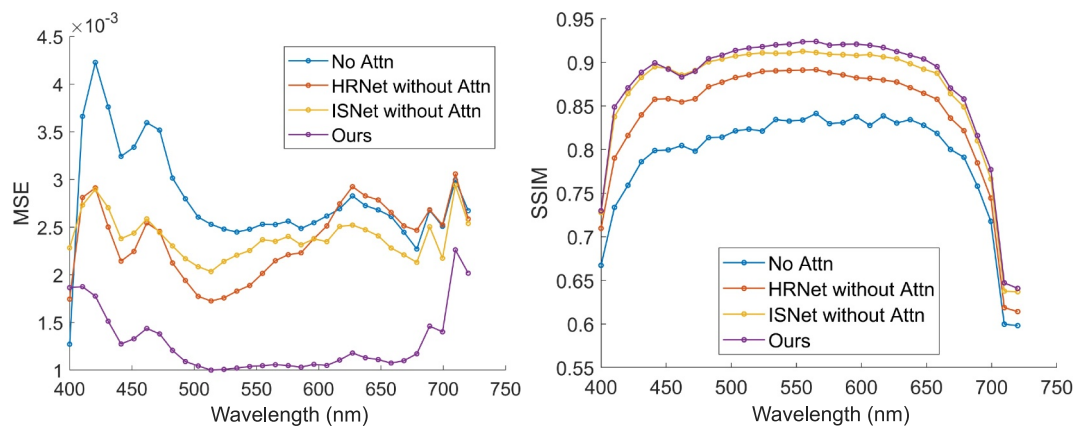


FIGURE 6 MSE (left) and SSIM (right) results of four methods computed from the entire 32 spectral bands. MSE, mean squared error; SSIM, structural similarity index measure.

imaging devices and controlled laboratory environments, thereby improving its feasibility in clinical practice.

We conduct extensive experiments on a self-constructed dataset comprising over 4000 RGB–hyperspectral image pairs to evaluate the effectiveness of the proposed framework. Experimental results indicate that the MSE value of the spectral reconstruction is 0.002, with the value of SSIM achieving 0.8496. Furthermore, ablation studies demonstrate the critical contribution of attention mechanisms to overall performance, as their

removal leads to degradation across all evaluation metrics.

Despite the promising results, the current framework is confined to dental scenarios. Future efforts will aim to generalize the approach to other materials and application domains, demonstrating its broader potential across different spectral reconstruction tasks. Given its scalability and data efficiency, the proposed inverse imaging framework also holds promise for broader biomedical applications, including tissue characterization and wound assessment.

4 | MATERIALS AND METHODS

4.1 | Principle of spectral imaging

Equation (1) presents a mathematical formula related to optical imaging, describing the imaging process from the tooth's spectral reflectance to the linear RGB image. In this equation, (x, y) represents a point on the tooth. $S(x, y, \lambda)$ denotes the spectral reflectance at point (x, y) , where λ is the spectral dimension. $I(\lambda)$ represents the spectral distribution function of the illumination source. $R_k(\lambda)$ represents the camera sensitivity curve for the k^{th} channel (specifically, when the imaging result is observed by the human eye, $R_k(\lambda)$ becomes the visual sensitivity for the k^{th} channel of the human eye). $k \in \{R, G, B\}$ refers to the spectral channel index, corresponding to the red, green, and blue channels, respectively.

$$G_k(x, y) = \int_{380 \text{ nm}}^{780 \text{ nm}} I(\lambda) S(x, y, \lambda) R_k(\lambda) d\lambda \quad (1)$$

This experimental scheme is derived from Equation (1), establishing Equations (2) and (3) to simulate the imaging processes of the restored tooth and the white board, respectively.

$$G_1(x, y) = \int_{380 \text{ nm}}^{780 \text{ nm}} I(\lambda) S(x, y, \lambda) C_k(\lambda) d\lambda \quad k \in \{R, G, B\} \quad (2)$$

$$G_2(x, y) = \int_{380 \text{ nm}}^{780 \text{ nm}} I(\lambda) S_0 C_k(\lambda) d\lambda \quad k \in \{R, G, B\} \quad (3)$$

Equation (2) describes the imaging process of the restored tooth. The RGB image of the restored tooth ($G_1(x, y)$) is obtained by integrating the spectral contribution across the entire visible wavelength range, where the light source spectrum ($I(\lambda)$) is modulated by the spectral reflectance of the restored tooth ($S(x, y, \lambda)$) and weighted by the camera's three-channel spectral sensitivity ($C_k(\lambda)$).

Equation (3) describes the imaging process of the white board. Based on actual measurements, the spectral reflectance of the white board is approximately 0.88 across the visible spectrum, and its RGB image ($G_2(x, y)$) is obtained by integrating the light source spectrum ($I(\lambda)$), weighted by the camera's three-channel spectral sensitivity ($C_k(\lambda)$) and the white board reflectance ($S_0 = 0.88$), over the visible wavelength range.

Although Equations (1–3) express the spectral imaging process as continuous integrals over the visible wavelength range [380 nm, 780 nm], actual imaging systems work with discrete sampled wavelengths due to sensor and filter limitations. Therefore, the integral equations are numerically approximated using discrete summations. Equations (2) and (3) can be expressed as Equations (4) and (5) by discretization, respectively.

$$G_1(x, y) = \sum_{i=1}^N I(x, y) S(x, y, \lambda_i) C_k(\lambda_i) \Delta\lambda \quad (4)$$

$$G_2(x, y) = \sum_{i=1}^N I(x, y) S_0 C_k(\lambda_i) \Delta\lambda \quad (5)$$

From the above equations, the transformation from hyperspectral data to RGB images can be interpreted as a process of spectral dimensionality reduction and color space mapping. However, reconstructing hyperspectral reflectance from RGB images is an ill-posed inverse problem due to the significant spectral dimensionality reduction in the forward imaging process.

To address the challenge, we propose to establish a mapping function from RGB images to hyperspectral reflectance data using a deep neural network, which captures the statistical patterns and structural distributions of dental color characteristics from a large number of training samples. As shown in Equation (6), the mapping function f is implemented as a neural network that estimates the hyperspectral reflectance with N_λ bands from an RGB image, effectively approximating the inverse of the forward imaging process.

$$f: \mathbb{R}^3 \rightarrow \mathbb{R}^{N_\lambda} \quad (6)$$

4.2 | Spectral reconstruction neural network

The reconstruction process consists of three interconnected neural networks: (1) a camera spectral sensitivity estimation network (the pre-trained MSNet); (2) an illumination spectrum estimation subnetwork; and (3) a hyperspectral reflectance reconstruction subnetwork. The subsequent sections provide a detailed description of these networks.

4.2.1 | Camera spectral sensitivity estimation network

Because the spectral reflectance of the tooth is device-dependent, to obtain accurate spectral reflectance, the

correct camera sensitivity function should be put into the network, helping the neural network learn spectral reflectance values closer to the ground truth. Based on our previous work of the camera spectral sensitivity estimation, the camera spectral sensitivity estimation network is applied to rebuild the camera spectral sensitivity from only a single ColorChecker image. The structure of this network is shown in Figure 7, with further details provided in Ref. [25].

4.2.2 | Illumination spectrum estimation network

As previously demonstrated, in addition to the spectral sensitivity of the camera, the spectral power distribution of the illumination is also a critical factor influencing the final RGB image. However, in many hyperspectral reconstruction tasks, the spectral distribution of the illumination used during image acquisition is often unknown. To address this issue, we design an illumination spectrum estimation network to predict the spectral power distribution of the light source, providing accurate prior illumination information for the entire framework. This helps mitigate the influence of illumination in the image, enabling the subsequent network modules to focus more effectively on recovering the intrinsic spectral reflectance of the restorative dental materials. The internal structure of the network is illustrated in Figure 8.

In this study, we propose the ISNet that takes the linear RGB image of a white board as the primary input, in conjunction with the camera spectral sensitivity, to predict the spectrum distribution of the illumination at the time of image acquisition. The estimated illumination spectrum serves as a critical prior input to the downstream reflectance reconstruction network.

The network adopts an encoder–decoder architecture to accomplish the spectrum estimation task. This architecture enables multi-scale feature extraction while preserving spatial information, effectively supporting the modeling of both local and global illumination-related features within the image. The encoder progressively extracts high-level semantic features, whereas the decoder, equipped with skip connections, reconstructs spatial details by fusing shallow and deep features. This ensures that the network maintains both semantic understanding and precise spatial localization.

To further enhance the network's performance in illumination estimation, we introduce an attention-based feature fusion module, incorporating both channel and SA mechanisms. The CA dynamically reweights spectral feature channels to emphasize those most relevant to accurate illumination prediction. Meanwhile, the SA mechanism helps identify illumination-dominant regions, thereby enabling more accurate modeling of the ambient spectral components.

4.2.3 | Hyperspectral reflectance reconstruction network

To recover the intrinsic spectral reflectance of tooth under specific lighting, we propose a spectral reflectance reconstruction network that removes devices' influence and isolates the material's true spectral characteristics. This module is essential for achieving physically interpretable and high-fidelity color reproduction. The internal structure of the network is illustrated in Figure 9.

In the network, the linear RGB image of the restored tooth is provided as the core input, guided by the estimated illumination spectrum and camera

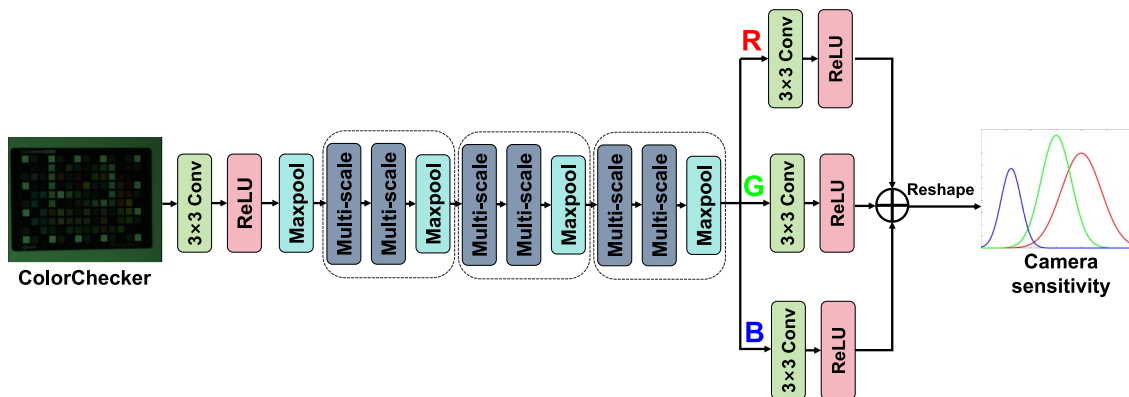


FIGURE 7 The structural details of camera spectral sensitivity network (the pre-trained MSNet). This network takes the ColorChecker's linear RGB image as input and utilizes multi-scale convolutional operations to extract the spectral sensitivity of the R, G, and B channels separately. These are then integrated to generate the final camera sensitivity.

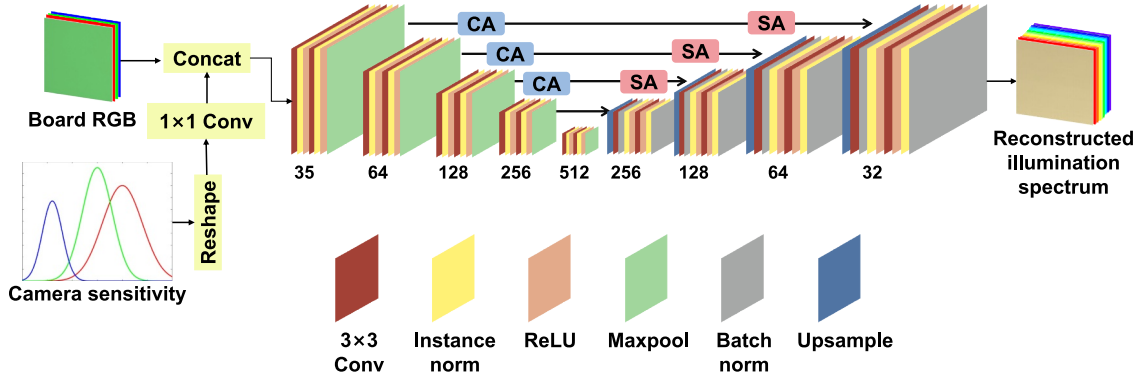


FIGURE 8 The structure of the illumination spectrum estimation network. The network takes the linear RGB image of a standard white board and the camera spectral sensitivity as input. The output is a reconstructed illumination spectrum.

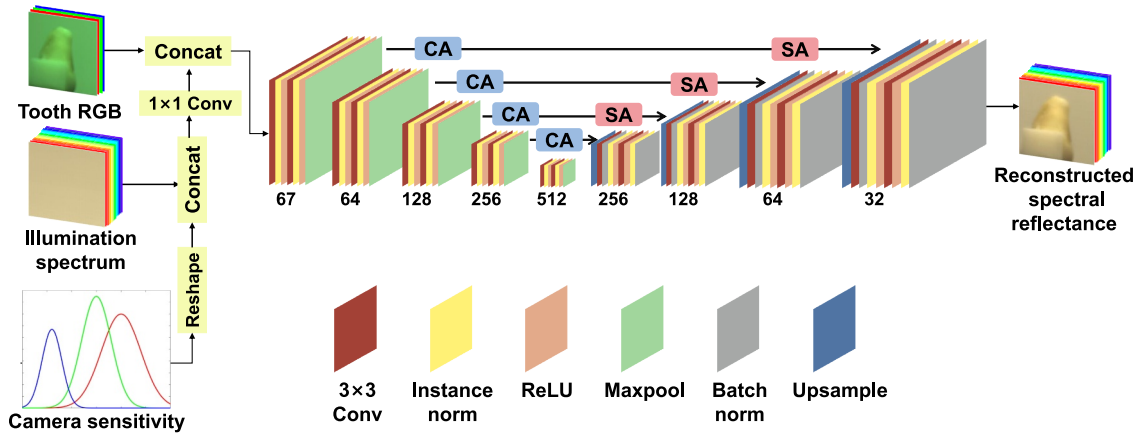


FIGURE 9 The structure of the hyperspectral reflectance reconstruction network. The network takes as input the linear RGB image of the tooth, the estimated illumination spectrum, and the camera spectral sensitivity. The output is a reconstructed hyperspectral reflectance cube.

spectral sensitivity, which serve as physical priors. These complementary components jointly guide the reconstruction of the intrinsic spectral reflectance. The RGB image provides perceptual cues of color and texture, which represents a complex mixture of illumination spectrum, tooth reflectance, and camera sensitivity. To supply physical context, we introduce the illumination spectrum as a prior that characterizes the illumination condition during image capture, allowing the network to discount lighting effects in the reconstruction process. Meanwhile, the camera spectral sensitivity enables the network to understand how the observed RGB values are derived from underlying spectral interactions. This design enhances the physical interpretability of the model and significantly improves its generalization capability under varying lighting conditions.

The network adopts the same encoder–decoder architecture as the ISNet, enabling effective extraction of multi-scale feature information. Through a series of convolutional layers and down-sampling operations, the network progressively compresses the input data to

capture both global and local features. Subsequently, up-sampling layers combined with skip connections are employed to recover the spatial resolution and achieve accurate reflectance prediction. To enhance feature representation, attention mechanisms, including both channel and SA modules, are incorporated to emphasize and refine features relevant to spectral reflectance reconstruction. Finally, a fully connected layer maps the decoder output to the predicted spectral reflectance. Given that the overall architecture is identical to that of the illumination estimation network, this section does not elaborate on the detailed structural design, and instead focuses on the fusion of multi-modal inputs and the physical interpretation of the reconstructed reflectance.

4.3 | Attention mechanism

To enhance feature representation and improve the model's ability to focus on the most informative spectral and spatial cues, we integrate attention mechanisms

into the network. Specifically, we employ CA and SA modules, which adaptively reweight features along different dimensions. The details of the CA and SA modules are described as follows.

The structure of the CA mechanism is shown in Figure 10. Given an input tensor $A \in \mathbb{R}^{W \times H \times C}$, the CA mechanism begins by applying convolutional operations to generate three feature matrices: the query matrix (Q), the key matrix (K), and the value matrix (V). These matrices are then reshaped and transposed to facilitate subsequent matrix multiplication. The reshaping operation aids in computing attention weights efficiently, whereas the transposition adjusts dimensional ordering to satisfy the requirements of matrix operations. Subsequently, Q and K are multiplied and weighted to obtain the channel weight. The channel weight is then applied to V through matrix multiplication. Finally, the resulting feature representation is fused with the original input to produce the output tensor $B \in \mathbb{R}^{W \times H \times C}$. This fusion al-

lows the learned attention weights to be integrated into the feature map while preserving the original spatial and spectral information, thereby enhancing the representational capacity of subsequent network layers.

The structure of the SA mechanism is shown in Figure 11. Using $C \in \mathbb{R}^{W \times H \times C}$ as input, convolution operations are first applied to obtain the three feature matrices: Q^* (query), K^* (key), and V^* (value). These matrices are then reshaped and transposed—reshaping facilitates matrix multiplication for attention weight computation, and transposition adjusts the dimension order to meet the requirements of subsequent operations. Next, Q^* and K^* are multiplied and weighted to obtain the spatial weight. Finally, the spatial weight is multiplied with V^* , and the result is fused with the original input to produce the output $D \in \mathbb{R}^{W \times H \times C}$. This fusion incorporates the SA weights into the features while preserving the original information for use in the following network layers.

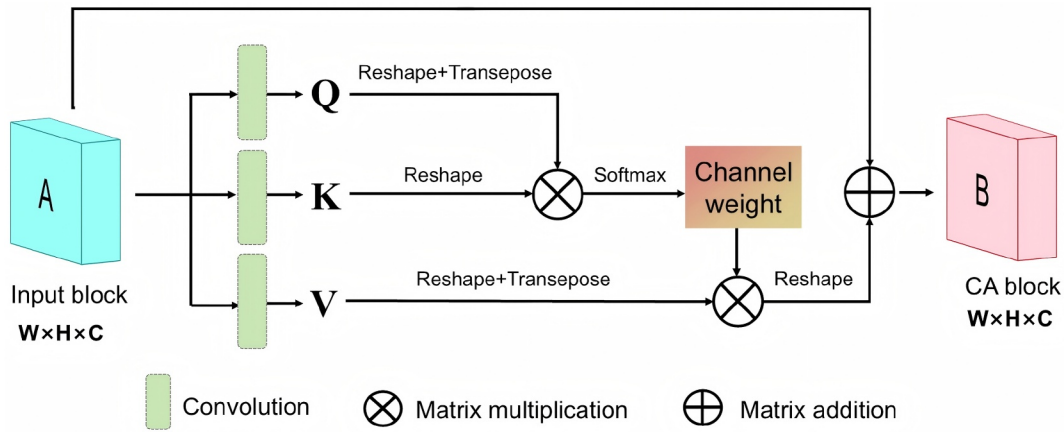


FIGURE 10 Structural details of the channel attention module.

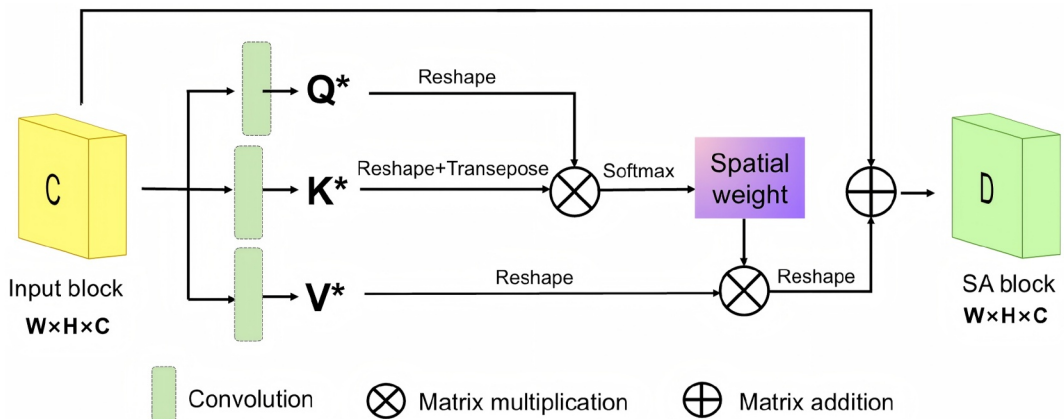


FIGURE 11 Structural details of the spatial attention module.

4.4 | Loss function

To jointly optimize the accuracy and structural fidelity of the reconstructed spectra and illumination, we design a composite loss function that integrates MSE, cosine similarity loss, and structural similarity loss. Specifically, the overall objective is defined in Equation (7):

$$L_{\text{total}} = (L_{\text{MSE}}^{(s)} + L_{\text{Cos}}^{(s)} + L_{\text{SSIM}}^{(s)}) + (L_{\text{MSE}}^{(w)} + L_{\text{Cos}}^{(w)} + L_{\text{SSIM}}^{(w)}) + \lambda L_{\text{reg}} \quad (7)$$

where the superscripts (s) and (w) denote spectral reflectance and light source estimation, respectively. L_{reg} is the L2 regularization term with weight λ . This combined loss encourages the network to achieve numerical accuracy, maintain spectral shape consistency, and preserve structural similarity under varying illumination conditions.

4.5 | Experimental setup

To achieve spectral-RGB image data acquisition, an optical system is constructed, as shown in Figure 12. In the experimental setup, the illumination provided by the lightbox irradiates the restored tooth and a white board. The light source interacts with the restored tooth, creating a coupling effect, whereas the white board performs diffuse reflection of the light source. After being split by a semi-transparent mirror, part of the light transmits into the center of the hyperspectral camera lens for imaging, whereas the other part reflects onto

the center of the RGB camera lens for imaging. By varying the spectral characteristics of the light source and capturing the resulting data from the restored tooth and the white board under different spectral light sources, the system's practical robustness and reliability can be enhanced.

4.5.1 | Light source system

Considering model stability and practical application scenarios, this experiment uses a self-constructed pseudo-parallel light source system [26], consisting of a lighting unit, diffusion structure, and control module. The system integrates 11 types of LEDs (3 white and 8 monochromatic) with a total of 1700 LEDs on a circuit board. Using custom software, the intensity of each LED can be controlled in real time to generate the required spectral light sources. For the capture of tooth and ceramic images, three broadband white LED illuminants with correlated color temperatures of 3000 K (warm white), 4000 K (neutral white), and 5600 K (cool white) were adopted, corresponding to typical indoor and daylight lighting conditions. The 5600 K source approximates the standard daylight illuminant (D65). In addition, monochromatic LEDs were employed to construct the spectral reflectance measurement environment.

4.5.2 | Acquisition system

The acquisition system consists of two parts: the hyperspectral acquisition system and the RGB image acquisition system. The hyperspectral acquisition system utilizes the hyperspectral camera. It collects spectral information of the tooth and white board within the wavelength range of 397.32–1003.58 nm, using 204 spectral bands, and records the data in RAW format. The RGB image acquisition system employs the RGB camera. The settings are adjusted to ensure that the captured data is not demosaicked, resulting in TIFF format of the captured images.

4.6 | Dataset preparation

4.6.1 | Sample preparation

To obtain a dataset aligned with the research objectives, we collaborated with West China Hospital of Sichuan University to prepare a set of dental samples with varying thickness gradients. This study is approved by the Medical Ethics Committee of West China Hospital of Sichuan University (No. WCHSIRB-CT-2022-257) and adheres to the medical ethics principles

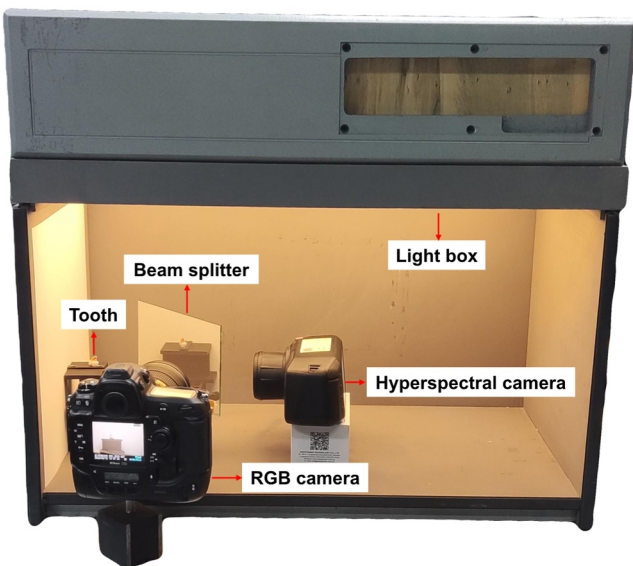


FIGURE 12 Image data acquisition optical system.

outlined in the Declaration of Helsinki (2013) and the International Ethical Guidelines for Health-Related Research Involving Humans by the Council for International Organizations of Medical Sciences (2016).

Natural teeth were obtained from patients at West China Hospital of Stomatology, Sichuan University. Extractions were performed between December 1 and 31 December, 2023, due to severe periodontitis or orthodontic treatment. Only premolars and molars without wear, caries, restorations, developmental anomalies, or severe discoloration were included. After cleaning and disinfection, the teeth were preserved in a 0.1% thymol solution. Restorative samples were fabricated by bonding custom zirconia ceramic slices to flattened natural tooth surfaces using dual-cure resin cement (U200, 3M), followed by light curing at 565 nm for 20 s. The zirconia slices were designed using CAD/CAM milling and subsequently sintered, ground, and polished to achieve a final thickness of 0.4–2.0 mm, covering the clinically relevant veneer range. In total, 10 restorative teeth were prepared for this study. These consisted of premolars and molars in approximately equal proportions.

4.6.2 | Image preprocessing

The data required for each group of training samples includes linear RGB images of the ColorChecker under the same light source, linear RGB images of the teeth, hyperspectral images of the teeth, linear RGB images of the white board, and hyperspectral images of the white board.

After demosaicing, the linear RGB images undergo two calibration steps: white balance correction to ensure accurate color representation under the acquisition lighting, and geometric calibration using a checkerboard pattern to align the RGB images with the hyperspectral images, compensating for differences in spatial resolution and field of view. The RGB images are then resampled using nearest neighbor interpolation to match the hyperspectral resolution, enabling

pixel-level correspondence for subsequent fusion or analysis. The hyperspectral data are additionally processed with visible-light band extraction. The final calibrated RGB–hyperspectral image pairs are presented in Figure 13.

To help the constructed network model better learn the data features and achieve spectral reconstruction, data augmentation is performed through random rotations, random horizontal shifts, random vertical shifts, and random shear transformations. Although using a constant value of 0 for padding, additive Gaussian noise with a zero mean and a standard deviation of 0.1 is added to a certain percentage of the data to enrich the dataset.

4.7 | Training details

Our framework is trained on 4400 sets (4000 for training and 400 for validation) and tested on 400 sets of images. Each data set comprises RGB captures of a ColorChecker, a restored tooth, and a white board, along with corresponding hyperspectral reflectance data for the tooth and the white board. All data were first acquired under identical acquisition conditions (illumination, camera, and geometry). After completing data collection, the dataset was randomly shuffled and divided into training, validation, and test subsets, ensuring that no sample appeared in more than one subset. Consequently, the test set is independent of the training and validation sets with respect to sample identity, while sharing the same acquisition environment.

The training process employed the Adam optimizer with a weight decay of $4e^{-4}$ and a learning rate of $1e^{-4}$. The batch size is set to 12, and the model is trained for 200 epochs. The training loss function is a weighted combination of MSE loss, SSIM-based structural loss, and cosine distance loss, applied separately to both the reflectance reconstruction and illumination estimation outputs. The total loss also included L2 regularization on network weights. Training and testing are conducted on Tesla P100 GPU using Tensorflow. During

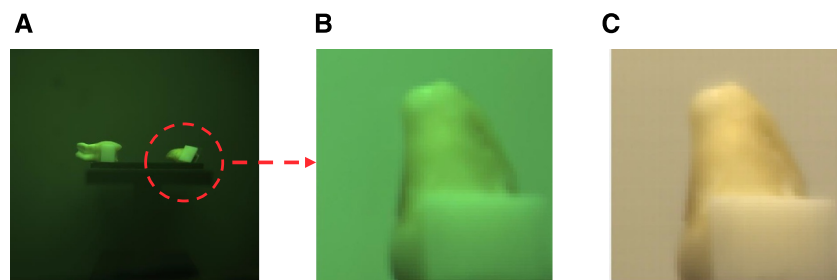


FIGURE 13 The final calibrated RGB–hyperspectral image pairs are presented. (A) The linear RGB image after demosaic algorithm. (B) The linear RGB image of the tooth after calibration (including white balance and geometric calibration). (C) The hyperspectral image of the tooth after registration for visualization.

inference, the model required an average of 0.95 s per batch, with a total runtime of 40.86 s for the entire test set, indicating potential feasibility for real-time clinical applications.

AUTHOR CONTRIBUTIONS

Yujun Feng: Formal analysis; investigation; methodology; validation; visualization; writing—original draft.

Yushan Guo: Data curation; formal analysis; investigation.

Qingyi Liu: Data curation; investigation; visualization.

Fei Liu: Funding acquisition; project administration; resources.

Jiefei Shen: Funding acquisition; resources; supervision.

Junfei Shen: Conceptualization; funding acquisition; methodology; project administration; resources; software; supervision; writing—review and editing.

ACKNOWLEDGMENTS

The authors would like to thank West China Hospital of Stomatology Sichuan University for providing the dental samples used in this study. This research was funded by the National Key Research and Development Program of China (2022YFC2410102); Sichuan Science and Technology Program (2025YFHZ0333); Research and Develop Program, West China Hospital, Sichuan University (RD-03-202408); Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) (2022QNRC001); Sichuan “Tianfu Emei Plan” talent project (A0103602).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

ETHICS STATEMENT

This study is approved by the Medical Ethics Committee of West China Hospital of Sichuan University (No. WCHSIRB-CT-2022-257) and adheres to the medical ethics principles outlined in the Declaration of Helsinki (2013) and the International Ethical Guidelines for Health-Related Research Involving Humans by the Council for International Organizations of Medical Sciences (2016).

REFERENCES

- [1] Cheng JX, Xie XS. Vibrational spectroscopic imaging of living systems: an emerging platform for biology and medicine. *Science*. 2015;350(6264):aaa8870.
- [2] Calin MA, Parasca SV, Savastu D, Manea D. Hyperspectral imaging in the medical field: present and future. *Appl Spectrosc Rev*. 2013;49(6):435–47.
- [3] Tsuda H, Arends J. Raman spectroscopy in dental research: a short review of recent studies. *Adv Dent Res*. 1997;11(4):539–47.
- [4] Zavala-Alonso V, Loyola-Rodríguez JP, Terrones H, Patiño-Marín N, Martínez-Castañón GA, Anusavice K. Analysis of the molecular structure of human enamel with fluorosis using micro-Raman spectroscopy. *J Oral Sci*. 2012;54(1):93–8.
- [5] He T, Zhang Q, Zhou M, Kou T, Shen J. Single-shot hyperspectral imaging based on dual attention neural network with multi-modal learning. *Opt Express*. 2022;30(6):9790–813.
- [6] Green RO, Eastwood ML, Sarture CM, Chrien TG, Aronsson M, Chippendale BJ, et al. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sens Environ*. 1998;65(3):227–48.
- [7] Huang L, Luo R, Liu X, Hao X. Spectral imaging with deep learning. *Light Sci Appl*. 2022;11(1):61.
- [8] Lawrence KC, Park B, Windham WR, Mao C. Calibration of a pushbroom hyperspectral imaging system for agricultural inspection. *Trans ASAE (Am Soc Agric Eng)*. 2003;46(2):513–21.
- [9] Lim HT, Murukeshan VM. Pushbroom hyperspectral imaging system with selectable region of interest for medical imaging. *J Biomed Opt*. 2015;20(4):046010.
- [10] Abdo M, Badilita V, Korvink J. Spatial scanning hyperspectral imaging combining a rotating slit with a dove prism. *Opt Express*. 2019;27(15):20290–304.
- [11] Kuznetsov K, Lambert R, Rey JF. Narrow-band imaging: potential and limitations. *Endoscopy*. 2006;38(1):76–81.
- [12] Guo Y, Song H, Liu H, Wei H, Yang P, Zhan S, et al. Model-based restoration of underwater spectral images captured with narrowband filters. *Opt Express*. 2016;24(12):13101–20.
- [13] Song H, Ma Y, Han Y, Shen W, Zhang W, Li Y, et al. Deep-learned broadband encoding stochastic filters for computational spectroscopic instruments. *Adv Theory Simul*. 2021;4(3):2000299.
- [14] Zhang W, Song H, He X, Huang L, Zhang X, Zheng J, et al. Deeply learned broadband encoding stochastic hyperspectral imaging. *Light Sci Appl*. 2021;10(1):108.
- [15] Thomas JB, Lapray PJ, Le Moan S. Trends in snapshot spectral imaging: systems, processing, and quality. *Sensors (Basel)*. 2025;25(3):675.
- [16] Shi Z, Dun X, Wei H, Dong S, Wang Z, Cheng X, et al. Learned multi-aperture color-coded optics for snapshot hyperspectral imaging. *ACM Trans Graph*. 2024;43(6):1–11.
- [17] Yuan X, Brady DJ, Katsaggelos AK. Snapshot compressive imaging: theory, algorithms, and applications. *IEEE Signal Process Mag*. 2021;38(2):65–88.
- [18] Ding K, Wang M, Chen M, Wang X, Ni K, Zhou Q, et al. Snapshot spectral imaging: from spatial-spectral mapping to metasurface-based imaging. *Nanophotonics*. 2024;13(8):1303–30.
- [19] Dai Q, Shimura T, Zheng Z. *Front Matter., Optoelectron. Imaging Multimed. Technol. X: SPIE*; 2023.12767.
- [20] Ahmed MT, Monjur O, Khaliduzzaman A, Kamruzzaman M. A comprehensive review of deep learning-based hyperspectral image reconstruction for agri-food quality appraisal. *Artif Intell Rev*. 2025;58(4):96.
- [21] Yu Z, Liu D, Cheng L, Meng Z, Zhao Z, Yuan X, et al. Deep learning enabled reflective coded aperture snapshot spectral imaging. *Opt Express*. 2022;30(26):46822–37.
- [22] Zhang B, Yuan X, Deng C, Zhang Z, Suo J, Dai Q. End-to-end snapshot compressed super-resolution imaging with deep optics. *Optica*. 2022;9(4):451–4.
- [23] Zheng S, Liu Y, Meng Z, Qiao M, Tong Z, Yang X, et al. Deep plug-and-play priors for spectral snapshot compressive imaging. *Photonics Res*. 2021;9(2):B18.

- [24] Hua X, Wang Y, Wang S, Zou X, Zhou Y, Li L, et al. Ultra-compact snapshot spectral light-field imaging. *Nat Commun.* 2022;13(1):2732.
- [25] Zhou M, Chen W, He T, Zhang Q, Shen J. Scan-free end-to-end new approach for snapshot camera spectral sensitivity estimation. *Opt Lett.* 2021;46(23):5806–9.
- [26] Shen J, Chang, Wang H, Zheng Z. Optimising the illumination spectrum for enhancing tissue visualisation. *Light Res Technol.* 2019;51(1):99–110.

How to cite this article: Feng Y, Guo Y, Liu Q, Liu F, Shen J, Shen J. Physics-informed multimodal learning for snapshot dental spectral reflectance prediction. *Quantitative Biology.* 2026;e70030. <https://doi.org/10.1002/qub2.70030>