International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-025-03525-8

ORIGINAL ARTICLE

Neural illumination calibration for surgical workflow-optimized spectral imaging

Alexander Baumann[1,2,7] · Leonardo Ayala[2,7] · Alexander Studier-Fischer[6,8,9,10] · Jan Sellner[2,3,4,5] · Berkin Özdemir[6,9,10] · Karl-Friedrich Kowalewski[8,9,10] · Slobodan Ilic[1] · Silvia Seidlitz[2,3,4,5] · Lena Maier-Hein[2,3,4,5,7]

Received: 4 April 2025 / Accepted: 17 September 2025

Abstract

Purpose Hyperspectral imaging (HSI) is emerging as a promising novel imaging modality with various potential surgical applications. Currently available cameras, however, suffer from poor integration into the clinical workflow because they require the lights to be switched off or the camera to be manually recalibrated as soon as lighting conditions change.

Methods We propose a novel learning-based approach to recalibration of hyperspectral cameras during surgery that predicts the corresponding white reference image from an uncalibrated hyperspectral input, enabling spatially resolved, automatic, and sterile calibration under varying illumination conditions. Our key novelty lies in (i) the disentanglement of the space of possible illuminations from the space of possible tissue configurations and (ii) combining real-world white reference measurements with physics-inspired simulated illuminations to create a diverse and representative training set.

Results Based on a total of 1,890 HSI cubes from a phantom, porcine subjects, rats, and humans, we derive the following key insights: Firstly, dynamically changing lighting conditions in the operating room dramatically reduce the performance of methods for physiological parameter estimation and surgical scene segmentation. Secondly, our method is not only sufficiently accurate to replace the tedious process of white reference-based recalibration, but also outperforms previously proposed methods by a large margin. Finally, our approach generalizes across species, lighting conditions, and image processing tasks.

Conclusion Our method enables seamless integration of hyperspectral imaging into surgical workflows by providing rapid and automated illumination calibration. Its robust generalization across diverse conditions significantly enhances the reliability and practicality of spectral imaging in clinical settings, paving the way for broader adoption of HSI in surgery.

Keywords Illumination Calibration · Hyperspectral Imaging · Intra-Operative Imaging · Deep Learning

Introduction

Hyperspectral imaging (HSI) has gained prominence in medical imaging, offering enhanced spectral information

Silvia Seidlitz and Lena Maier-Hein contributed equally to this work.

✉ Alexander Baumann
baumann.alexander@siemens.com

1 Siemens AG, Munich, Germany

2 Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

3 Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany

4 Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany

5 National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and university medical center Heidelberg, Heidelberg, Germany

6 Department of General, Visceral, and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany

7 Medical Faculty, Heidelberg University, Heidelberg, Germany

8 Department of Urology and Urosurgery, University Medical Center Mannheim, Mannheim, Germany

9 Division of Intelligent Systems and Robotics in Urology (ISRU), German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

10 DKFZ Hector Cancer Institute, University Medical Center Mannheim, Mannheim, Germany

Published online: 07 October 2025

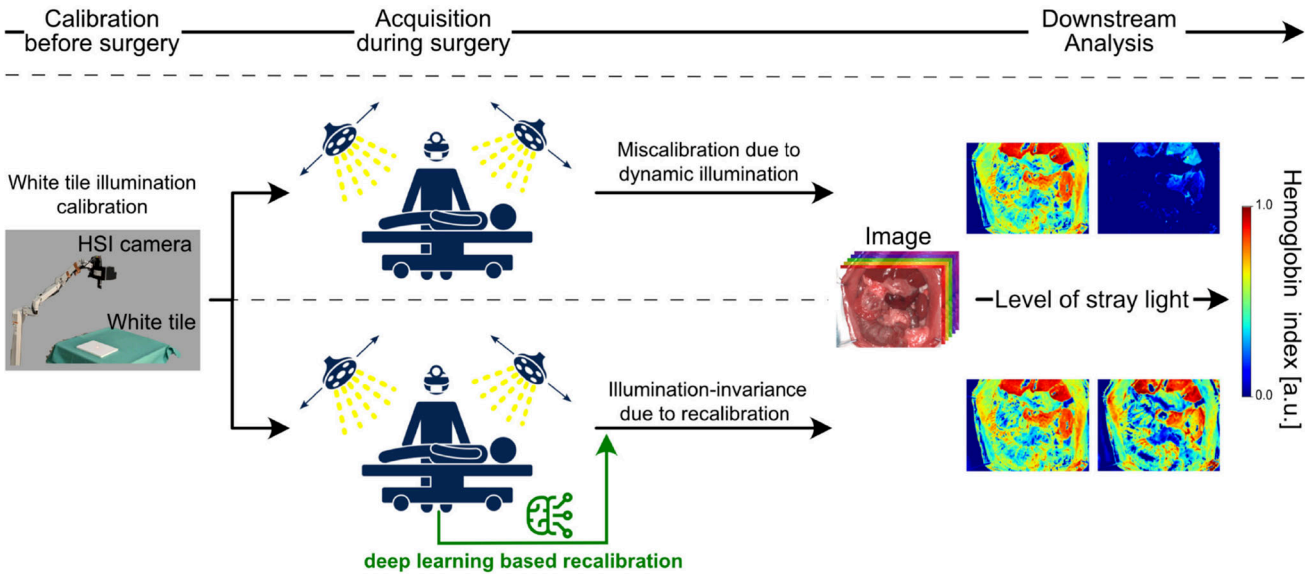

Fig. 1 Current hyperspectral cameras, reliant on static illumination environments, fail in real-world scenarios with dynamically changing lighting conditions

over conventional RGB imaging. Recent studies highlight its benefits in tissue classification [1–5] and its potential for estimating physiological tissue parameters [6–12]. However, in open surgery, spectral data are susceptible to variations in illumination, necessitating proper calibration whenever lighting conditions change [13]. Current experimental protocols require switching off all external light sources for accurate data capture [14]. Given that these measures considerably disrupt the clinical workflow, it is presumed not to be consistently applied, leading to compromised data integrity and degraded downstream task performance, as illustrated in Fig. 1. This may contribute to the limited clinical adoption of spectral imaging.

Conventional HSI light calibration utilizes a physical white tile measurement, representing the surrounding illumination. This method, however, is constrained by temporal and sterility factors, rendering it impractical in the operating room (OR) context. While the proposed use of white sterile OR rulers [15] addresses sterility concerns, it introduces challenges due to inaccuracies and an increased workload. Several automatic calibration algorithms, originally developed for RGB imaging, can be adapted for HSI, recovering a global scene illuminant based on intensity statistics [16–18]. Alternatively, HSI-specific calibration leverages specular highlights for illuminant estimation using similar statistical methods [19]. However, learning-based methods have demonstrated superior performance over their non-learning counterparts for RGB calibration [20, 21]. Recently a deep learning approach tailored for multispectral illumination calibration was developed [22]. Adapting an RGB-based technique [20], this method processes multispectral channel triplets in the log-chroma space using convolutional operations. However, applying this approach to hyperspectral images results in high computational costs due to the exponential increase in channel triplet combinations. More critically, all preceding methods assume spatially uniform illumination—an unrealistic simplification for OR environments [23]. In RGB imaging, multi-illuminant color constancy models have shown promise in overcoming this challenge, predicting pixel-wise illuminants for calibration [24–26]. In spectral imaging, however, only one deep learning approach has thus far addressed multi-illuminant calibration [27]. Specifically, Li et al. framed illumination calibration as a factorization problem between surface reflectance and illumination intensity, proposing an optimization framework based on unrolling networks to extract the surface reflectance information.

Overall, the methods proposed in the literature either remain untested for surgical HSI, which presents illumination characteristics distinct from natural scenes [23], and/or are conceptually not suitable for spatially resolved calibration. Given this bottleneck, the mission of our work was to develop a new workflow-optimized calibration approach that enables widespread clinical spectral imaging. Our specific contribution is fourfold:

1. *Limitations of current methods:* We demonstrate that dynamically changing lighting conditions in the OR dramatically affect the performance of in vivo HSI applications, and previously proposed calibration methods fail to restore optimal performance.
2. *Need for spatially resolved calibration:* We show that spatially uniform illuminants cannot adequately represent the illumination within the OR, underlining the

necessity of spatially resolved illumination calibration methods.
3. *New learning-based approach:* We present a novel learning-based approach to performing spatially resolved light recalibration of surgical hyperspectral images. Specifically, we propose to replace conventional physical white reference measurements with a data-driven prediction of the corresponding white tile measurement. This enables a seamless and sterile recalibration process during surgery.
4. *Large-scale validation across three species:* Based on the downstream tasks of semantic segmentation and physiological parameter estimation, we show that our recalibration method not only outperforms previous methods, but also generalizes across species, lighting conditions, and image processing tasks.

With this contribution, we substantially expand our preliminary version published at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) [28] by (i) validating the benefit of our method on a human patient cohort with a total of 1,148 hyperspectral images, (ii) investigating the additional research question on spatially resolved illuminants, (iii) performing comprehensive and fine-grained analyses of core design choices, and (iv) including two additional baseline methods.

Materials and methods

The primary challenge in data-driven calibration lies in generalizing to unseen settings To render our method conceptually robust to domain shifts, we propose estimating the white tile measurement corresponding to a given scene rather than directly predicting the recalibrated image. The underlying premise of our methodology is that the acquisition of a comprehensive training dataset, encompassing all potential tissue and illumination configurations, is practically unattainable. We therefore disentangle the space of possible illuminations from the space of possible tissue configurations, as illustrated in Fig. 2. Specifically, a two-dataset training paradigm is implemented for the neural network. The first dataset, referred to as the illumination dataset, comprises a collection of real and simulated white tile images, capturing diverse lighting conditions prevalent within the OR. The second dataset consists of accurately calibrated images of clinically relevant samples. To simulate a stray light-affected HSI cube, an image in the sample dataset is augmented through element-wise multiplication with a white reference image. Subsequently, the neural network is trained to reconstruct the illumination from the input. During inference, an uncalibrated hyperspectral image, possibly contaminated by stray light, is fed into the neural network to predict the white reference image required for illumination calibration.

To contextualize the design choices of our model, we first provide a concise theoretical overview of how light influences image formation, based on [29], before elaborating on our illumination generation strategy and neural network architecture.

Theoretical background

Assuming negligible specular reflections, image formation can be modeled as integral of the product of surface reflectance R, illumination intensity L, and the camera's spectral sensitivity S, which determines the captured wavelength range for each channel. Specifically, an image I is modeled as:

$$I_c(x, y) = \int_{\Lambda_c} R(x, y, \lambda) L(x, y, \lambda) S_c(\lambda) d\lambda \tag{1}$$

where (x, y) denotes the spatial coordinates, c the channel index, and $\Lambda_c = \{\lambda \in \mathbb{R} | S_c(\lambda) \neq 0\}$ represents the support of S_c. Consequently, a white reference image, I_{white}, representing the surrounding illumination, enables calibration through element-wise division, as expressed by $I_{\text{cal}} = I \oslash I_{\text{white}}$. Synthetic relighting of a calibrated scene is then accomplished by element-wise multiplication with an arbitrary illuminant.

Datasets

Model training and validation were performed exclusively on porcine data, whereas testing systematically assessed stray light conditions on unseen porcine subjects, a phantom, rats, and humans, as summarized in Fig. 3. While the phantom colorchecker board dataset was acquired with the Tivita® 2.0 Surgery (Diaspective Vision GmbH, Am Salzhaff, Germany) featuring light-emitting diode (LED) illumination, the others were captured with the halogen-based Tivita® Tissue. As these light sources exhibit different behavior when interfering with the main stray light source, namely LED-based surgical lights [23], we evaluated on both systems.

Measured illumination dataset: The purpose of this dataset was to acquire a variety of representative OR lighting conditions for algorithm training. To this end, white reference images were acquired within an OR, accounting for camera light and various stray light sources, including Dr. Mach LED surgical lights (Model: 8MC), ceiling lights, and daylight. Diverse stray light scenarios were achieved by varying the angle, distance, and number of surgical lights as well as adjusting blinds or ceiling light, resulting in a wide range of illumination spectra. As the two HSI systems in this study differ in the light sources, we captured one illumina-

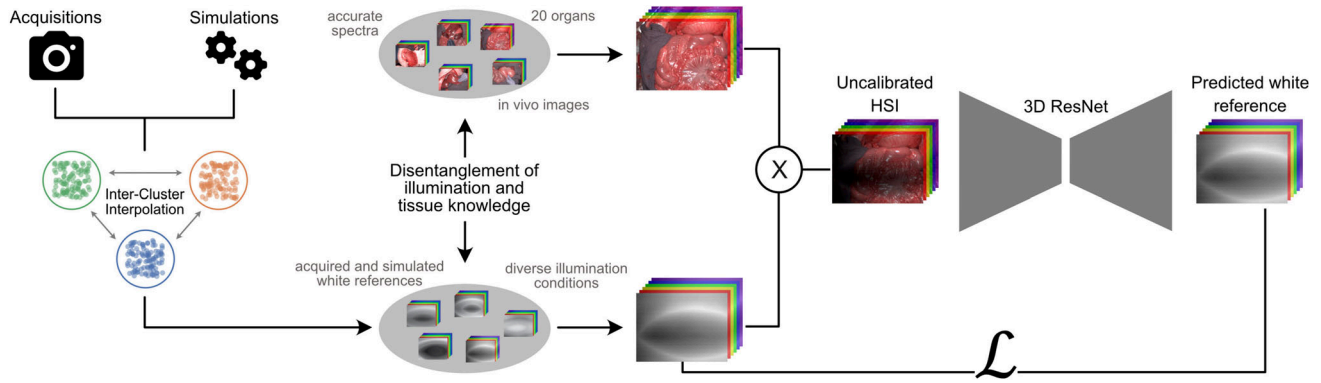

Fig. 2 The proposed approach substitutes labor-intensive manual calibration with a dynamic fully-automated approach. At the heart of our data-centric method lies a 3D convolutional neural network, trained on in vivo data augmented with synthetic illumination variations. At inference, the model processes a raw hyperspectral image and generates the corresponding white reference image. The resulting white tile prediction facilitates the subsequent calibration of the input image

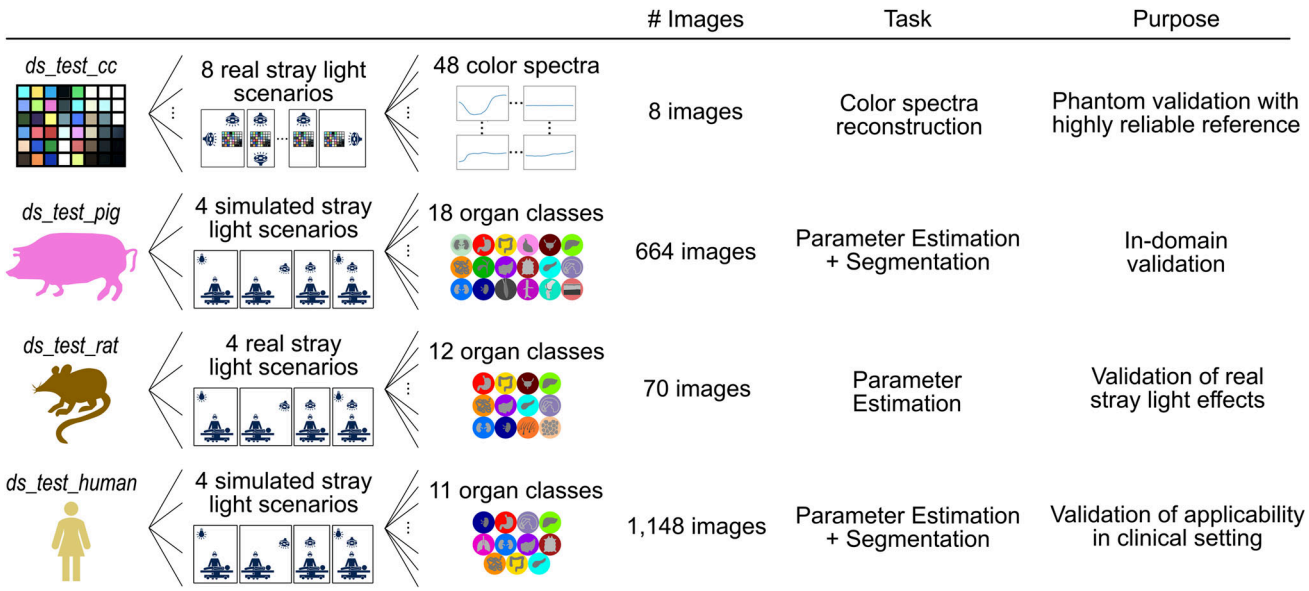

Fig. 3 Testing concept based on data from a phantom and three species

tion dataset with each camera: *ds_dev_ill_led* (LED) and *ds_dev_ill_hal* (halogen).

In vivo porcine development dataset: For model development, we curated the publicly available HSI dataset HeiPorSPECTRAL [14], partitioning it into a training set, *ds_dev_pig*, and a validation set, *ds_val_pig*. Notably, these datasets exhibit no overlap with the downstream datasets used for testing.

Test datasets: The proposed methodology underwent thorough evaluation utilizing the four datasets *ds_test_cc*, *ds_test_pig*, *ds_test_rat*, and *ds_test_human*—detailed in Fig. 3. To evaluate the model's calibration efficacy, the evaluation strategy comprised both simulated and real stray light scenarios, as recommended by [29–31]. Colorchecker boards acquired under diverse lighting conditions assessed recalibration performance based on high-fidelity reference data (*ds_test_cc*). In-domain testing was performed on *ds_test_pig*, while the model's generalizability and clinical applicability were evaluated on *ds_test_rat* and *ds_test_human*.

Both *ds_test_pig* and *ds_test_human* were acquired under controlled, constant lighting conditions to ensure accurate calibration; additional details can be found in [1, 32]. To simulate stray light effects in these datasets, we generated an illumination dataset, *ds_test_ill_hal*, comprising white reference images captured under four distinct lighting setups. These setups vary in both the intensity and position of the stray light source: (1) The first scenario uses only ceiling light as stray light, (2) the second consists of a Dr. Mach LED surgical light illuminating the surgical field from the right,

(3) the third positions the surgical light to shine directly onto the surgical site, and (4) the fourth combines both ceiling light and direct surgical light illumination on the site.

In contrast, the rat organ images were acquired under real stray light conditions. Each scene was first captured without stray light and subsequently re-imaged under the four illumination setups described above. For each illumination configuration, two acquisitions were made, resulting in a total of 10 images per subject. The complete process for acquiring all lighting setups per subject took approximately 4.5 min, during which we ensured the rat's physiological stability and maintained a consistent scene.

Physics-based illumination simulation

To implement the data-centric recalibration concept, we focused on the model-based generation of plausible white tile data. To overcome the resource-intensive acquisitions, we enhanced both $ds\_dev\_ill\_led$ and $ds\_dev\_ill\_hal$ by synthesizing illumination images from real white tile images. Since surgical overhead lights, predominantly LED-based, are the main source of stray light in ORs [23] and induce distinct interference patterns in our LED and halogen-based HSI systems, we developed tailored simulation strategies for each. While LED-based simulations comprise interpolations of entire hyperspectral images, our halogen-based strategy first simulates one-dimensional light spectra and subsequently synthesizes full hyperspectral images by incorporating realistic spatial variations in intensity.

LED simulations: For the LED-based HSI system, wave interference with LED-based surgical lights is approximately constructive; thus, local extrema in the spectrum of the camera light source are preserved. To model this interference behavior, we employed an efficient interpolation strategy. Initially, we grouped existing white tile images into $K = 4$ clusters through the K-means algorithm, with each cluster representing a distinct illumination scenario. This clustering was performed on spatially averaged spectra, providing sparse representations of the illuminations. To create new white tile images, we then conducted inter-cluster interpolation by combining randomly selected images from different clusters.

Halogen simulations: Halogen-based HSI systems, unlike their LED-based counterparts, experience destructive wave interference when used alongside LED surgical lights due to differences in local extrema. Consequently, standard interpolation techniques, which inherently preserve local extrema, prove inadequate. To address this limitation, we introduce a simulation strategy that can shift extrema positions, effectively accommodating this spectral variability. The complementary roles of interpolations and simulations are illustrated in Suppl. Figure 1. Our simulation approach synthesizes hyperspectral images by first modeling one-dimensional light spectra using a parametric function and subsequently including realistic spatial variations in intensity. Inspired by Planck's radiation law and empirical observations, we define the parametric function as follows:

$$f_{p_1,p_2,p_3,p_4}(\lambda) = \frac{(p_1\lambda - p_2)^3}{\exp(p_3\lambda - p_4) - 1} \tag{2}$$

where f_{p_1,p_2,p_3,p_4} is the intensity, λ is the wavelength and $p_1, \ldots, p_4$ are the parameters. For each instance of $ds\_dev\_ill\_hal$, least-squares optimization was conducted to determine parameters that best approximate the given instance. Subsequently, the mean μ_k and standard deviation σ_k of parameter p_k were calculated across all instances. Using a fixed number of standard deviations, n_σ, intervals P_k for parameter p_k were defined as $P_k = [\mu_k - n_\sigma\sigma_k; \mu_k + n_\sigma\sigma_k]$. Subsequently, sampling $(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4)$ from the grid $P_1 \times P_2 \times P_3 \times P_4$ yields a simulated light spectrum $f_{\tilde{p}_1,\tilde{p}_2,\tilde{p}_3,\tilde{p}_4}$. Illumination images were then synthesized using these light spectra as the spatially averaged spectrum. Specifically, a hyperspectral image I was randomly selected from $ds\_dev\_ill\_hal$ and normalized by its mean spectrum. To generate the simulated white reference image I_s, the normalized image was element-wise multiplied with the simulated spectrum $f_{\tilde{p}_1,\tilde{p}_2,\tilde{p}_3,\tilde{p}_4}$, thereby preserving the original spatial variations. Concretely, for spatial indices (i, j), and the channel index c, we calculate:

$$I_s(i, j, c) = f_{\tilde{p}_1,\tilde{p}_2,\tilde{p}_3,\tilde{p}_4}(\lambda_c) \odot I(i, j, c) \oslash \overline{I}(c) \tag{3}$$

where λ_c denotes the wavelength associated to channel c, and $\overline{I}$ represents the spatially averaged spectrum of image I. To further enhance the coverage of illumination conditions, inter-cluster interpolations based on the acquisitions and simulations were generated, as conducted for LED simulations.

Neural network training details

As illustrated in Fig. 2, illumination images are multiplied with stray light-free sample images to simulate uncalibrated hyperspectral images under varying lighting conditions. To generate spectrally smooth white reference predictions, a 3D convolutional neural network (CNN) is employed. The network utilizes an autoencoder architecture incorporating ResNet blocks [33] in both the encoder and decoder. Two design choices were critical for the efficacy of the presented method. Firstly, during training, the model is optimized solely on the predicted white reference image, rather than the resulting calibrated sample image, thereby reducing the reliance on the sample dataset. Particularly, the mean-squared error between the predicted and original white reference image serves as the loss function $\mathcal{L}$. Secondly, conventional U-Net-style skip connections between the encoder

and decoder are intentionally omitted to minimize the influence of geometric sample information within the predicted illumination image. With these design choices in place, the encoder gradually enlarges the feature dimension from 32 to a 256-dimensional latent space across nine ResNet blocks, using strided convolutions for down-sampling at every third block. The decoder mirrors this process, restoring the original resolution via interpolation. All convolutional layers, except for the input layer, employ a 3×3 kernel. The model comprises approximately 13.5 million trainable parameters. Optimization was performed using the Adam optimizer with an exponentially decaying learning rate scheduler.

Experiments and results

We investigated the following research questions (RQs):

(RQ1) How do dynamically changing lighting conditions in the OR affect the performance of hyperspectral image analysis?
(RQ2) Is the common practice of using spatially uniform lighting models adequate for capturing the complexity of OR illumination in dynamic recalibration?
(RQ3) Are neural networks capable of replacing white tile recalibration of hyperspectral cameras in the OR?
(RQ4) To what extent can neural network-based recalibration mitigate the performance drop of hyperspectral image analysis under varying lighting conditions?

Experiment RQ1

As baseline methods, we used the traditional approaches Gray world [16], Max-RGB [17], first-order Gray edge [18], and an HSI-specific method leveraging specular highlights [19]. Additionally, we incorporated two learning-based calibration techniques: AngularGAN [24], designed for RGB imagery, and an approach [27] tailored for multispectral data, which utilizes unrolling networks to extract surface reflectance. Both learning-based methods were trained on the porcine sample dataset ds_dev_pig, paired with the acquired illumination images. Semantic organ segmentation and physiological parameter estimation were conducted on in vivo data as downstream evaluation tasks. Stray light-affected versions of the ds_test_pig and ds_test_human organ datasets were created using the illumination test set $ds\_test\_ill\_hal$, wherein each version corresponded to a distinct illumination configuration of $ds\_test\_ill\_hal$. Subsequently, these images were recalibrated by one of the evaluated methods, followed by inference for the respective downstream task. For segmentation, we employed pig and human organ models sourced from [2, 32], which had been trained on accurately calibrated images. As a segmentation metric, the Dice similarity coefficient (DSC) was used [34] and compared against the performance on the dataset devoid of stray light. In the second downstream analysis, organ-specific physiological parameters, including oxygen saturation, perfusion, hemoglobin, and water index, were calculated according to the procedures outlined in [9]. Reference parameters were obtained by applying the methods to images acquired without stray light. Subsequently, the parameters were computed on the recalibrated images that were originally affected by stray light. Calibration performance was evaluated by the mean absolute error between the recalibrated parameters and their corresponding reference values. For both downstream tasks, the hierarchical structure of the data was respected during aggregation following [1].

The necessity of recalibration is qualitatively demonstrated in Fig. 1 and quantitatively supported by the substantial variance in Fig. 4 when recalibration is omitted. However, existing HSI calibration techniques lacked adequate accuracy. Even the best performing baseline method, AngularGAN, resulted in an average decrease of 21 % in DSC when applied to human data. Similar failures in downstream tasks could be observed on pig and rat images, as demonstrated in Suppl. Figure 3.

Experiment RQ2

To investigate whether spatially uniform illuminants are sufficient for calibration, we first isolated the problem from specific calibration methods. Particularly, images in ds_test_pig were relit by multiplying them with white reference images of $ds\_test\_ill\_hal$. Recalibration was then performed using the spatial average of the illumination image as a spatially uniform illuminant estimate. The impact of this reduction was evaluated through organ segmentation, with results compared against reference values obtained without stray light. The left plot of Fig. 5 illustrates the DSC relative to the average spatial standard deviation of the white reference images for the four illumination scenarios in $ds\_test\_ill\_hal$. A clear inverse correlation was observed, with DSC decreasing as the spatial deviation of the corresponding white tile image increased. Crucially, the sole limitation to spatially uniform illuminants resulted in a DSC degradation of more than 9 % in certain lighting conditions. Analogous behavior was observed for oxygen saturation estimation, where the error proportionally increased with the standard deviation (see Suppl. Figure 2).

Furthermore, we performed an ablation study on the illuminant dimensionality within our proposed model architecture by training a comparable model that generates spatially uniform illuminants. For this, our model's encoder was combined with a pooling operation in the latent space, reducing spatial dimensions, and the decoder was adapted to use 1D convolutional blocks. The training strategy remained similar,

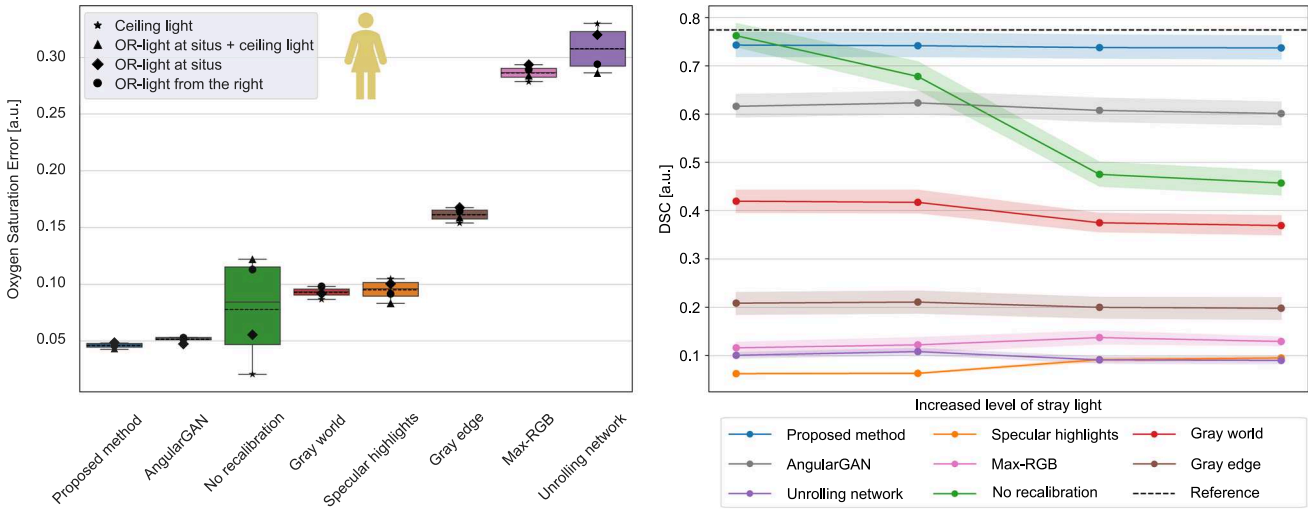

Fig. 4 Unlike existing calibration methods, our method produces accurate recalibrations, ensuring robust downstream capabilities on human data. (Left) Absolute errors in organ-specific oxygen saturation estimates between reference images and recalibrated images. (Right) Segmentation results on the recalibrated images. Shaded regions: 95 % confidence intervals

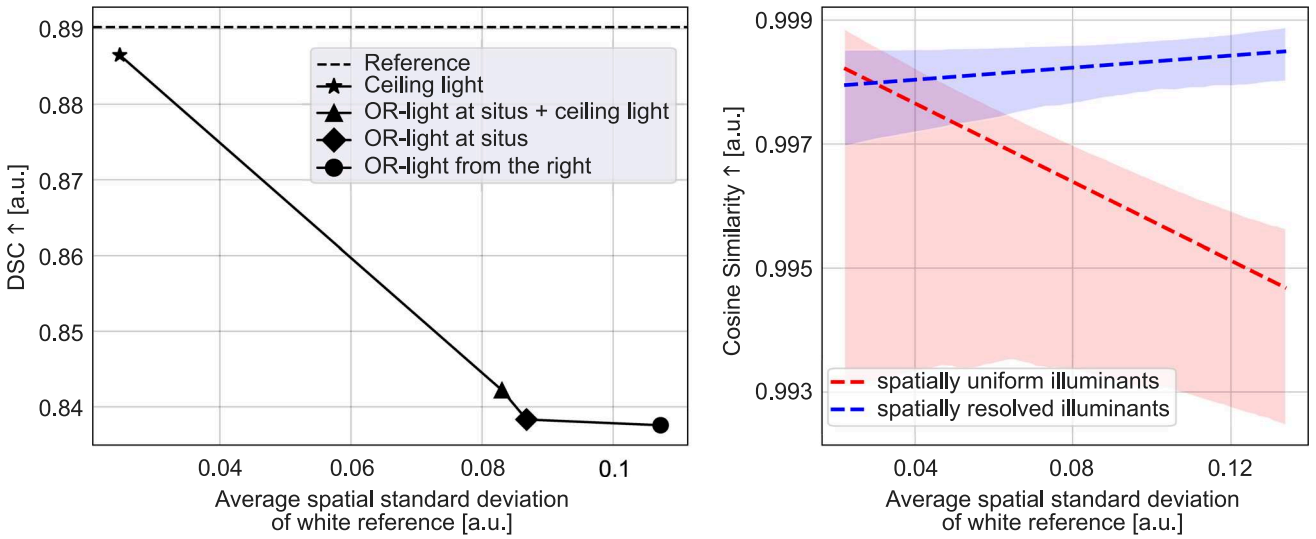

Fig. 5 Spatially uniform illuminants cannot adequately represent the surrounding illumination within the OR. Calibration accuracy (right) and downstream performance (left) decrease due to the limitation to spatially uniform illuminants. (Left) Organ segmentation on images that were reilluminated with white reference images and recalibrated with the respective spatially averaged spectrum as spatially uniform illuminant. (Right) Cosine similarities between reference images and images recalibrated using the spatially uniform model (red) and the spatially resolved model (blue). Dashed lines: fitted regression lines; shaded regions: 95 % confidence intervals

except that the illumination images were spatially averaged. For validation, images from *ds_val_pig* were relit by white reference images and subsequently recalibrated using both the spatially uniform model and the proposed spatially resolved model. To gauge calibration performance, the cosine similarity between recalibrated and original images was calculated and illustrated on the right of Fig. 5. Consistent with previous observations, calibration accuracy of the spatially uniform model decreased with increasing spatial standard deviation of the illumination images. In contrast, the spatially resolved model demonstrated stable—if not slightly increased—cosine similarity scores. Although the absolute differences between models appear modest, a statistical t-test confirmed the superior performance of the spatially resolved model ($p = 2.57 \cdot 10^{-8}$). Across all ablation studies, cosine similarity scores remained generally high, resulting in only marginal differences between models. However, we observed that even subtle changes in cosine similarity can

Table 1 Our simulation strategy effectively enhances the illumination training data. As validation metric, the spectral cosine similarity between the original and recalibrated images was used. The pig organ validation dataset served as our sample data, augmented with white tile images, which were excluded from downstream analysis. In brackets: 95 % confidence intervals

Illumination Data for Training	Cosine Similarity ↑ (×100)
Acquisitions	99.75
	[99.69; 99.80]
+ Simulations	99.79
	[99.76; 99.82]

Table 2 Realism-diversity trade-off for simulated illuminations. The sampling range of simulation parameters is varied. $n_\sigma = 0.25$ yields the highest cosine similarity, indicating an optimal balance. In brackets: 95 % confidence intervals

Standard Deviations of Parameter Range	Cosine Similarity ↑ (×100)
$n_\sigma = 0.1$	99.78
	[99.75; 99.80]
$n_\sigma = 0.25$	99.79
	[99.76; 99.82]
$n_\sigma = 1$	99.69
	[99.62; 99.76]

produce pronounced effects on downstream tasks. Therefore, we recommend using this metric primarily for relative model ranking rather than for interpreting absolute performance values.

Experiment RQ3

The model's illumination calibration capabilities were evaluated using a downstream-invariant approach across in-distribution and out-of-distribution settings. In-distribution performance was examined by analyzing latent features generated by our model. To achieve this, illumination conditions of $ds\_test\_ill\_hal$ were simulated within the images of ds_test_pig. Subsequently, recalibration was performed using our trained model. Finally, the original, stray light-affected, and recalibrated images were encoded into the latent space via our trained encoder, and the first two PCA components were visualized using kernel density estimation. This visualization was conducted for each stray light scenario, as illustrated in Fig. 6. The model's ability to differentiate between the original and stray light-affected images correlated with the level of stray light within the illumination images, suggesting meaningful feature extraction. Furthermore, the distribution of our recalibrations closely resembled the original distribution, indicating high calibration accuracy.

To further assess the calibration accuracy in an out-of-distribution setting with a highly reliable reference, we recalibrated the colorchecker board dataset ds_test_cc using our model. As illustrated in Suppl. Figure 2, the proposed method achieved the highest average cosine similarity (0.9905) among the baseline methods from Section 3.1, closely approaching the gold standard white tile calibration (0.9945).

The model's design choices, particularly the impact of halogen-based illumination simulations, were analyzed via an ablation study on the validation set ds_val_pig. Table 1 shows that incorporating simulations to real data improved performance, demonstrating the effectiveness of our simulation strategy. This benefit also translated to downstream applications, where segmentation performance increased by +0.05 DSC on the validation set. To determine the hyperparameter n_σ that controls widths of the parameter ranges, we tested three values. As can be seen in Table 2, large values of n_σ led to unrealistic simulations and performance drops. Finally, the architectural choice of omitting skip connections is analyzed in Suppl. Table 1.

Experiment RQ4

To assess the downstream capability of our method, we replicated the experiments outlined in Section 3.1, incorporating our recalibration approach. Following the findings presented in Table 2, the hyperparameter n_σ was set to 0.25 for all experiments, which were carried out on untouched sample and illumination test sets. As shown in Figs. 4 and 7, and in Suppl. Figure 3, our method outperformed previous methods by a large margin. In particular, the proposed approach yielded high segmentation scores on pig and human organ images across diverse illumination setups. Furthermore, the model excelled in the parameter estimation task, consistently ranking first on pig and rat datasets compared to the baseline methods. This was further supported by the lowest average absolute error in oxygen saturation, shown in Fig. 4, and the stable hemoglobin estimates across real stray light scenarios, illustrated in Fig. 1.

Discussion

Our study is the first to provide in vivo evidence that dynamic illumination changes in the OR can lead to severe failures in HSI downstream analyses. The clinical implications are substantial, as manual recalibration considerably impedes the efficiency of surgical workflows and may limit the widespread adoption of HSI cameras in clinical practice. Notably, the proposed method stands as the only calibration method in our study that consistently achieves high accuracy regardless of the downstream task and domain, thereby

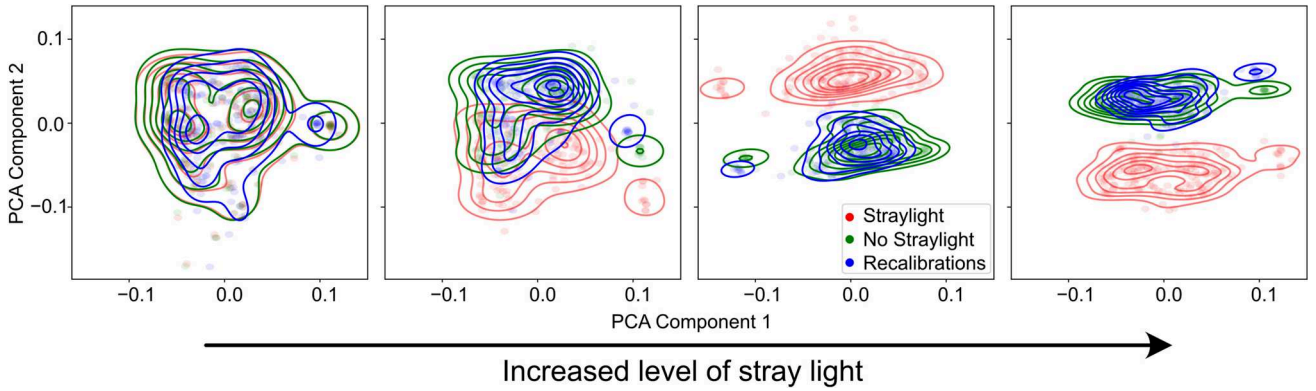

Fig. 6 The proposed model produces recalibrations (blue) from stray light-affected images (red) that resemble corresponding non-stray light-affected images (green) within the latent space. The features were extracted using the model's encoder, pooled into 1D feature vectors, and then subjected to PCA

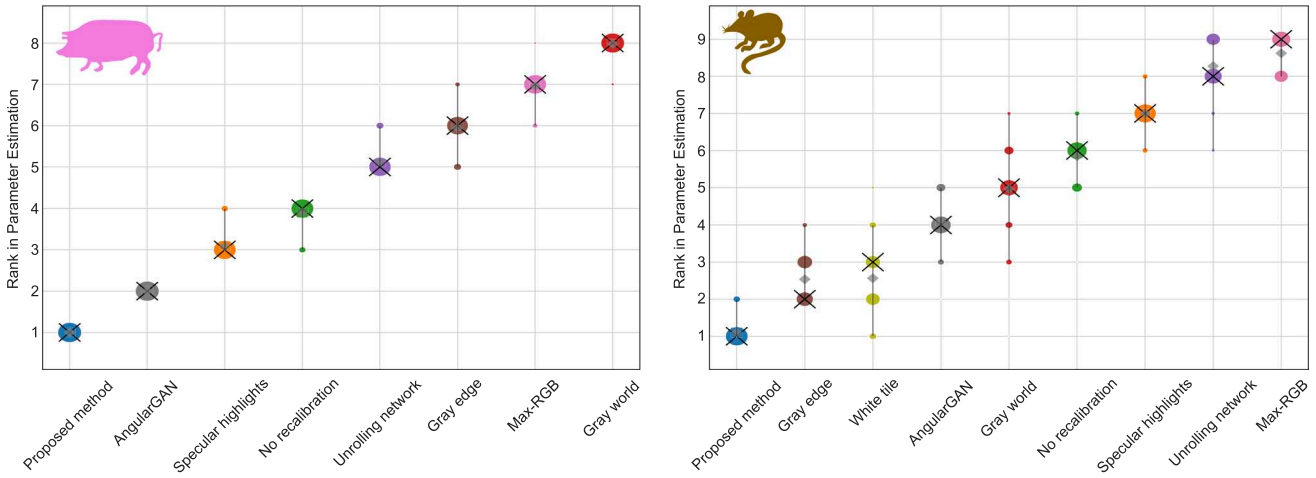

Fig. 7 Our model showcases the highest robustness against simulated and real stray light interference in the estimation of physiological parameters. The accompanying plot evaluates calibration models on pig (left) and rat (right) organ images by ranking them according to the absolute error of four physiological parameters. Each blob's area reflects the relative frequency at which the rank is achieved across 1000 bootstrap samples, following [35]

highlighting its high clinical applicability. It further offers several key conceptual benefits. First, white reference measurements are not only impractical due to sterilization and workflow constraints, but they are also susceptible to oversaturation, which accounts for the suboptimal performance observed on the rat data (see Fig. 7). In contrast to the majority of existing methods, our approach accommodates spatial deviations in illumination, thereby enabling enhanced calibration accuracy (see Fig. 5). A key strength of the presented approach is its inherent generalizability. It exhibits superior performance to the competing neural network method, AngularGAN, even under identical training conditions—a performance gap we attribute to the intrinsic domain shift in auto-encoded images. Consequently, our model's ability to seamlessly handle human data without sacrificing calibration accuracy reinforces its applicability in clinical settings (see Fig. 4). Additionally, our algorithm is well-suited for practical deployment, processing each image in approximately 0.71 s of which only 0.06 s is required for model inference on a NVIDIA TITAN RTX GPU. With a modest GPU memory footprint of about 1.4 GB, integration into existing surgical HSI systems is both feasible and efficient.

While our work may be limited by not addressing every conceivable illumination scenario, we prioritized key light sources prevalent in ORs and validated our model on highly diverse datasets. Consequently, we are confident that our conclusions will hold in a wide range of settings.

In conclusion, this study introduces a novel learning-based illumination calibration method for spectral imaging. Our methodology not only delivers superior performance compared to existing approaches across diverse settings but also lends itself to seamless integration into HSI systems for ORs.

This advancement could thus facilitate the emergence of surgical workflow-optimized spectral imaging.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-025-03525-8.

Acknowledgements Funding Open Access funding enabled and organized by Projekt DEAL. This project was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (NEURAL SPICING, 101002198), the National Center for Tumor Diseases (NCT), Heidelberg's Surgical Oncology Program, the German Cancer Research Center (DKFZ), and the Helmholtz Association under the joint research school HIDSS4Health (Helmholtz Information and Data Science School for Health). We also acknowledge the support through state funds for the Innovation Campus Health + Life Science Alliance Heidelberg Mannheim from the structured postdoc program for Alexander Studier-Fischer: Artificial Intelligence in Health (AIH).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval The ethics for the animal experiments was granted by the Committee on Animal Experimentation of the Baden-Württemberg Regional Council in Karlsruhe, Germany (G-261/19, G-262/19, G-62/23). The HSI human data were obtained during the SPACE trial (SPectrAl Characterization of organs and tissuEs during surgery) at Heidelberg University Hospital, with approval from the Ethics Committee of the Medical Faculty of Heidelberg University, Germany (S-459/2020). The trial adhered to the ethical principles of the Declaration of Helsinki and the principles of Good Clinical Practice.

References

1. Seidlitz S, Sellner J, Odenthal J, Özdemir B, Studier-Fischer A, Knödler S, Ayala L, Adler TJ, Kenngott HG, Tizabi M (2022) Robust deep learning-based semantic organ segmentation in hyperspectral images. Med Image Anal 80:102488
2. Sellner J, Seidlitz S, Studier-Fischer A, Motta A, Özdemir B, Müller-Stich BP, Nickel F, Maier-Hein L (2023) Semantic segmentation of surgical hyperspectral images under geometric domain shifts. In: International Conference on Medical Image Computing and Computer- Assisted Intervention, pp. 618–627 (2023). Springer
3. Trajanovski S, Shan C, Weijtmans PJ, Koning SGB, Ruers TJ (2020) Tongue tumor detection in hyperspectral images using deep learning semantic segmentation. IEEE Trans Biomed Eng 68(4):1330–1340
4. Halicek M, Fabelo H, Ortega S, Callico GM, Fei B (2019) In-vivo and ex-vivo tissue analysis through hyperspectral imaging techniques: revealing the invisible features of cancer. Cancers 11(6):756
5. Halicek M, Lu G, Little JV, Wang X, Patel M, Griffith CC, El-Deiry MW, Chen AY, Fei B (2017) Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. J Biomed Opt 22(6):060503–060503
6. Clancy NT, Jones G, Maier-Hein L, Elson DS, Stoyanov D (2020) Surgical spectral imaging. Med Image Anal 63:101699
7. Ayala LA, Wirkert SJ, Gröhl J, Herrera MA, Hernandez-Aguilera A, Vemuri A, Santos E, Maier-Hein L (2019) Live monitoring of haemodynamic changes with multispectral image analysis. In: OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging: Second International Workshop, OR 2.0 2019, and Second International Workshop, MLCN 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 2, pp. 38–46. Springer
8. Ayala L, Adler TJ, Seidlitz S, Wirkert S, Engels C, Seitel A, Sellner J, Aksenov A, Bodenbach M, Bader P (2023) Spectral imaging enables contrast agent-free real-time ischemia monitoring in laparoscopic surgery. Sci Adv 9(10):6778
9. Kulcke A, Holmer A, Wahl P, Siemers F, Wild T, Daeschlein G (2018) A compact hyperspectral camera for measurement of perfusion parameters in medicine. Biomedical Engineering/Biomedizinische Technik 63(5), 519–527
10. Holmer A, Tetschke F, Marotz J, Malberg H, Markgraf W, Thiele C, Kulcke A (2016) Oxygenation and perfusion monitoring with a hyperspectral camera system for chemical based tissue analysis of skin and organs. Physiol Meas 37(11):2064
11. Wirkert SJ, Vemuri AS, Kenngott HG, Moccia S, Götz M, Mayer BF, Maier-Hein KH, Elson DS, Maier-Hein L (2017) Physiological parameter estimation from multispectral images unleashed. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20, pp. 134–141. Springer
12. Shapey J, Xie Y, Nabavi E, Bradford R, Saeed SR, Ourselin S, Vercauteren T (2019) Intraoperative multispectral and hyperspectral label-free imaging: a systematic review of in vivo clinical studies. J Biophotonics 12(9):201800455
13. Ebner M, Nabavi E, Shapey J, Xie Y, Liebmann F, Spirig JM, Hoch A, Farshad M, Saeed SR, Bradford R (2021) Intraoperative hyperspectral label-free imaging: from system design to first-in-patient translation. J Phys D Appl Phys 54(29):294003
14. Studier-Fischer A, Seidlitz S, Sellner J, Bressan M, Özdemir B, Ayala L, Odenthal J, Knoedler S, Kowalewski K-F, Haney CM (2023) Heiporspectral-the heidelberg porcine hyperspectral imaging dataset of 20 physiological organs. Sci Data 10(1):414
15. Bahl A, Horgan CC, Janatka M, MacCormac OJ, Noonan P, Xie Y, Qiu J, Cavalcanti N, Fürnstahl P, Ebner M (2023) Synthetic white balancing for intra-operative hyperspectral imaging. J Med Imag 10(4):046001–046001
16. Buchsbaum G (1980) A spatial processor model for object colour perception. J Franklin Inst 310(1):1–26
17. Land EH (1977) The retinex theory of color vision. Sci Am 237(6):108–129
18. Weijer J, Gevers T, Gijsenij A (2007) Edge-based color constancy. IEEE Trans Image Process 16(9):2207–2214
19. Ayala L, Seidlitz S, Vemuri A, Wirkert SJ, Kirchner T, Adler TJ, Engels C, Teber D, Maier-Hein L (2020) Light source calibration

for multispectral imaging in surgery. Int J Comput Assist Radiol Surg 15:1117–1125
20. Barron JT (2015) Convolutional color constancy. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 379–387
21. Hu Y, Wang B, Lin S (2017) Fc4: Fully convolutional color constancy with confidence-weighted pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4085–4094
22. Glatt O, Ater Y, Kim W-S, Werman S, Berby O, Zini Y, Zelinger S, Lee S, Choi H, Soloveichik E (2024) Beyond rgb: a real world dataset for multispectral imaging in mobile devices. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4344–4354
23. Sharma N, Heer A, Su L (2023) A timeline of surgical lighting-is automated lighting the future? The Surgeon 21(6):369–374
24. Sidorov O (2019) Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0
25. Hussain MA, Akbari AS (2018) Color constancy algorithm for mixed-illuminant scene images. IEEE Access 6:8964–8976
26. Das P, Liu Y, Karaoglu S, Gevers T (2021) Generative models for multi-illumination color constancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1194–1203
27. Li Y, Fu Q, Heidrich W (2021) Multispectral illumination estimation using deep unrolling network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2672–2681
28. Baumann A, Ayala L, Studier-Fischer A, Sellner J, Özdemir B, Kowalewski K-F, Ilic S, Seidlitz S, Maier-Hein L (2024) Deep intra-operative illumination calibration of hyperspectral cameras. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 120–131. Springer
29. Gijsenij A, Gevers T, Weijer J (2011) Computational color constancy: survey and experiments. IEEE Trans Image Process 20(9):2475–2489
30. Barnard K, Cardei V, Funt B (2002) A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data. IEEE Trans Image Process 11(9):972–984
31. Barnard K, Martin L, Coath A, Funt B (2002) A comparison of computational color constancy algorithms. ii. experiments with image data. IEEE Trans Image Process 11(9):985–996
32. Sellner J, Studier-Fischer A, Qasim AB, Seidlitz S, Schreck N, Tizabi M, Wiesenfarth M, Kopp-Schneider A, Knödler S, Haney CM et al (2024) Xeno-learning: knowledge transfer across species in deep learning-based spectral image analysis. arXiv preprint arXiv:2410.19789
33. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778
34. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, Glocker B, Isensee F, Kleesiek J, Kozubek M et al (2024) Metrics reloaded: recommendations for image analysis validation. Nature methods, 1–18
35. Wiesenfarth M, Reinke A, Landman BA, Eisenmann M, Saiz LA, Cardoso MJ, Maier-Hein L, Kopp-Schneider A (2021) Methods and open-source toolkit for analyzing and visualizing challenge results. Sci Rep 11(1):2369