PhySpec: Physically Consistent Spectral Reconstruction via Orthogonal Subspace Decomposition and Self-Supervised Meta-Auxiliary Learning

Xingxing Yang [1] **Jie Chen** [1] **Zaifeng Yang** [2]

Abstract

This paper presents a novel approach to hyperspectral image (HSI) reconstruction from RGB images, addressing fundamental limitations in existing learning-based methods from a physical perspective. We discuss and aim to address the "***colorimetric dilemma***": failure to consistently reproduce ground-truth RGB from predicted HSI, thereby compromising physical integrity and reliability in practical applications. To tackle this issue, we propose **PhySpec**, a physically consistent framework for robust HSI reconstruction. Our approach fundamentally exploits the intrinsic physical relationship between HSIs and corresponding RGBs by employing orthogonal subspace decomposition, which enables explicit estimation of camera spectral sensitivity (CSS). This ensures that our reconstructed spectra align with well-established physical principles, enhancing their reliability and fidelity. Moreover, to efficiently use internal information from test samples, we propose a self-supervised meta-auxiliary learning (MAXL) strategy that rapidly adapts the trained parameters to unseen samples using only a few gradient descent steps at test time, while simultaneously constraining the generated HSIs to accurately recover ground-truth RGB values. Thus, MAXL reinforces the physical integrity of the reconstruction process. Extensive qualitative and quantitative evaluations validate the efficacy of our proposed framework, showing superior performance compared to SOTA methods.

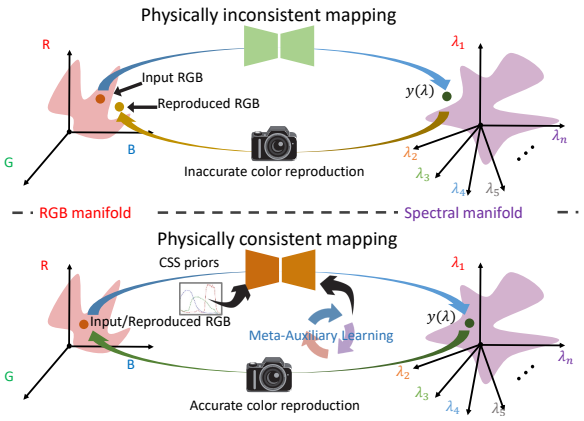

Figure 1. Physically inconsistent reconstruction (top, *i.e.*, most existing methods) and physically consistent reconstruction with physical constraints (camera spectral sensitivity (CSS)) and meta-auxiliary learning (bottom, *i.e.*, our method).

1. Introduction

"*We need to build in priors about the structure of the world, about physics, about causality, about the fact that the world is three-dimensional.* —— *Yann LeCun, 2018.* "

Hyperspectral images (HSIs) encode detailed radiance spectra for each pixel, offering significantly richer spectral information compared to traditional RGB images, which are limited to three bands including red, green, and blue. The spectral reflectance captured by HSIs reveals the intrinsic material properties of objects and remains unaffected by varying lighting conditions, which has been widely used in remote sensing (Borengasser et al., 2007; Yuan et al., 2017), medical imaging (Johnson et al., 2007; Lu & Fei, 2014), and scene relighting (Lam & Sato, 2013).

In computer vision, rapid developments in deep learning have paved an alternative way for hyperspectral image acquisition from RGB images in a data-driven learning manner (Shi et al., 2018; Huang et al., 2024; Yang et al., 2024). However, it is a challenging inverse problem to expect RGB images of just three channels to recover more than three degrees of freedom in spectral data (*e.g.*, 31 or even more than 100). Fortunately, a substantial portion of spectral variation

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China [2]Institute of High Performance Computing, Agency for Science Technology and Research, Singapore. Correspondence to: Jie Chen <chenjie@comp.hkbu.edu.hk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

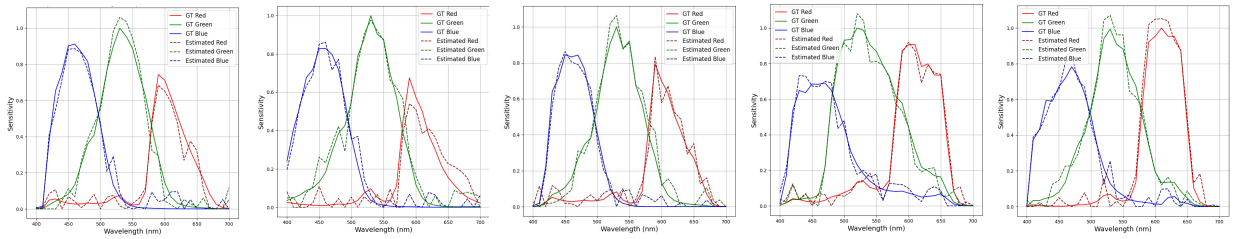

Figure 2. Gallery of our estimated camera spectral sensitivity compared with the ground-truth data.

is encapsulated in color appearance in natural scenes (*i.e.*, the RGB values), which allows learning-based methods to produce relatively accurate spectral approximations.

A straightforward question arises: ***Can existing methods really generate high-fidelity and reliable spectra?***

Our analysis reveals two core limitations in current learning-based approaches: **i, Physical Consistency Failure.** Reconstructed spectra must reproduce ground-truth RGB colors via CSS (camera spectral sensitivity) matching, mirroring the forward process of RGB capture. Yet, most state-of-the-art methods suffer from a ***colormetric dillema***: their predicted spectra fail to accurately reconstruct RGB values, exposing fundamental physical inconsistencies that undermine reliability (Fig. 1). **ii, Illumination-Dependent Generalization.** While spectral reflectance is illumination-invariant, RGB values are not. Models trained on fixed illumination conditions struggle with unseen data due to varying lighting, as they apply static parameters equally to all test samples (Yang et al., 2025). Zero-shot learning (Liu et al., 2018; Socher et al., 2013) partially addresses this by leveraging internal patterns within individual test images. However, single-image optimization often lacks robustness, limiting reconstruction accuracy.

To address these issues, we propose ***PhySpec***, a physically consistent spectral reconstruction framework that bridges data-driven learning and imaging physics. Firstly, to address the first limitation, we estimate CSS explicitly and apply orthogonal subspace decomposition, which is a well-studied method in inverse problems (Lin & Finlayson, 2020; Wang et al., 2023), to integrate physical constraints between spectra and RGB values, instead of relying on black-box learning. Unlike prior work (*e.g.*, Lin et al. (Lin & Finlayson, 2020)), which studies synthetic data with known CSS and enforces RGB reproduction within a single network (causing training instability), we further introduce **meta-auxiliary learning (MAXL)**, which assigns reconstruction as the primary task and reproduction as an auxiliary task in a self-supervised manner. Specifically, we first train both tasks simultaneously utilizing paired data (*i.e.*, external information). Second, we fine-tune the pre-trained parameters using each testing sample (internal information) while focusing on the auxiliary task in a self-supervised manner, which also addresses the second limitation of existing methods. In addition, we design a **dynamic illumination estimation module (DIEM)** to implicitly estimate an image-specific illumination descriptor without requiring prior knowledge of illumination conditions, which further addresses the second limitation of the existing methods and enhances generalization.

The main contributions are summarized as follows:

- We explicitly estimate the camera spectral sensitivity (CSS) and introduce orthogonal subspace decomposition to integrate intrinsic physical constraints between spectra and RGB values, formulating a physically consistent framework to alleviate the ***colormetric dillema*** that challenges existing models;
- We present the first framework that introduces self-supervised meta-auxiliary learning (MAXL) for spectral reconstruction, which enforces generated HSIs to accurately recover ground-truth RGBs, thereby ensuring physical integrity for the inverse problem;
- We design a dynamic illumination estimation module (DIEM) to estimate image-specific illumination descriptors implicitly with unknown illuminations, enhancing the generalization to practical applications;
- Extensive experiments demonstrate that PhySpec significantly outperforms SOTA methods for HSI reconstruction based on RGB inputs.

2. Related Works

Spectral Reconstruction. Early attempts relied on model-based frameworks that combined fidelity terms and physical priors to constrain solutions. For instance, Robles et al. (Robles-Kelly, 2015) incorporated color and texture prior to refine reconstruction. Building on this work, Arad et al. (Arad & Ben-Shahar, 2016) proposed the use of spectral priors to develop sparse dictionaries linking HSIs with their RGB counterparts. Nonetheless, these model-based methods rely on a fixed formulation based on several strong assumptions, which restricts their ability to adapt to the diverse and complex imaging conditions in various real-

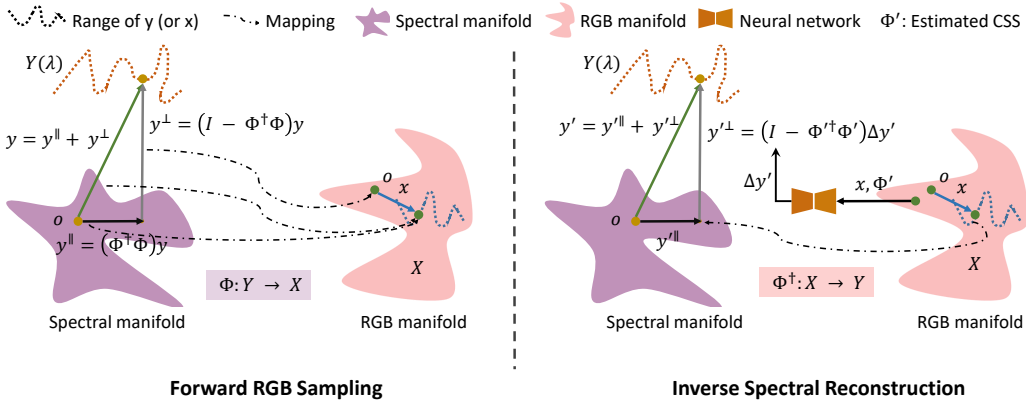

Figure 3. The schematic diagram of Orthogonal Subspace Decomposition. According to the forward RGB sampling process $\boldsymbol{x} = \boldsymbol{\Phi y}$, the spectra $\boldsymbol{y}$ can be uniquely decomposed into orthogonal subspaces: the range-space component $\boldsymbol{y}^{\parallel}$ and null-space $\boldsymbol{y}^{\perp}$. In the reconstruction phase, the object of the range-space component $\boldsymbol{y}'^{\parallel}$ can be directly accessed using the pseudo-inverse matrix $\boldsymbol{\Phi}^{\dagger}$ corresponding to $\boldsymbol{\Phi}$. While the other object of the null-space component $\boldsymbol{y}'^{\perp}$ is to first generate a raw prediction of spectral signal $\Delta\boldsymbol{y}'$ satisfying the physical constraints of explicitly estimated camera spectral sensitivities, and then obtained by the null-space projection operator $\boldsymbol{I} - \boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi}$.

world scenarios. In contrast, learning-based methods (Cai et al., 2022b; Yang et al., 2024; Zhu et al., 2021; Li et al., 2023a) have adopted data-driven strategies to infer implicit mappings from RGB to hyperspectral domains through specialized neural architectures. A breakthrough was achieved with the HSCNN (Xiong et al., 2017) model, which employed convolutional layers and deep residual blocks to transform RGB inputs into enriched HSI feature spaces. Further advancements were introduced by MST++ (Cai et al., 2022b), which leveraged a transformer-based framework with channel-wise self-attention to capture long-range spectral correlations and thus significantly improved the performance. However, these data-driven learning methods often overlook the physical consistency, leading to unreliable spectral predictions.

Meta-auxiliary Learning. MAXL (Liu et al., 2019) was initially proposed to enhance the generalization of classification models using meta-learning (Hospedales et al., 2021) to identify optimal labels for auxiliary tasks without requiring manually labeled auxiliary data. Chi et al. (Chi et al., 2021) applied MAXL to low-level vision tasks by integrating external and internal learning, designing a self-supervised auxiliary reconstruction task that partially shares the network with the primary deblurring task. This approach enables fast adaptation at test time, addressing distribution shifts identified by Sun et al. (Sun et al., 2020). Cheng et al. (Cheng et al., 2024) proposed a meta-transfer learning framework that incorporated multiple HSI datasets to address the data shortage dilemma for HSI super-resolution. Zhang et al. (Zhang et al., 2024) proposed an unsupervised test-time adaptation learning (UTAL) framework that jointly estimates unknown degradation and adapts a pre-trained deep image prior to specific hyperspectral images at test time, enabling effective super-resolution under complex real-world conditions. Huo et al. (Huo et al., 2024) employ MAXL for spectral reflectance recovery trained on synthetic RGB images. Their real RGB-HSI pairs combine different cameras that are aligned manually, but separate capture precludes pixel-wise pairing, risking inaccuracies and constrained generalization across device variations. In this work, we adopt the strategy of incorporating a self-supervised RGB recovery from predicted spectra as the auxiliary task to enforce the generated HSI to accurately reconstruct the corresponding ground-truth RGBs, preserving physical consistency and ensuring the integrity of the inverse problem.

3. Method

3.1. Preliminaries: Orthogonal Subspace Decomposition

Given a non-zero linear matrix $\mathcal{H} \in \mathbb{R}^{n\times m}$, a pseudo-inverse $\mathcal{H}^{\dagger} \in \mathbb{R}^{m\times n}$ usually holds that satisfies $\mathcal{H}\mathcal{H}^{\dagger}\mathcal{H} = \mathcal{H}$. Singular value decomposition (SVD) (Klema & Laub, 1980) can be used to compute the analytical solution by:

$$\mathcal{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}, \quad \mathcal{H}^{\dagger} = \mathbf{V}\mathbf{\Sigma}^{\dagger}\mathbf{U}^{\top}, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{\Sigma}$ is a diagonal matrix with eigenvalues as its diagonal elements.

Suppose $\Phi = \mathcal{H}^{\dagger}\mathcal{H}$ is a mapping between a linear space that can be seen as the operator that projects samples to the range-space of $\mathcal{H}$ since $\mathcal{H}\mathcal{H}^{\dagger}\mathcal{H} \equiv \mathcal{H}$. While $\Phi^{\dagger} = (\mathbf{I} - \Phi$ can be seen as the orthogonal operator that projects samples to the null-space of $\mathcal{H}$, since $\mathcal{H}\Phi^{\dagger} = \mathcal{H}(\mathbf{I} - \mathcal{H}^{\dagger}\mathcal{H}) \equiv \mathbf{0}$, where $\mathbf{I}$ is a unit matrix.

Any sample vector $\boldsymbol{x} \in \mathbb{R}^{m}$ can be expressed as the sum of two components: one that resides in the range space of $\mathcal{H}$

and another that lies in the null space of $\mathcal{H}$:

$$\boldsymbol{x} \equiv \Phi \boldsymbol{x} + \Phi^{\dagger} \boldsymbol{x}. \tag{2}$$

3.2. Problem Definition

The forward RGB rendering process can be simulated by calculating the inner products between the measured radiance spectra $\mathbf{Y} \in \mathbb{R}^{\lambda \times H \times W}$, illumination spectrum $\mathbf{L} \in \mathbb{R}^{\lambda}$, and the spectral sensitivities $\mathbf{S} \in \mathbb{R}^{3 \times \lambda}$ of a given RGB camera:

$$\mathbf{X}^{k}(u, v) = \int_{\lambda} \mathbf{S}^{k}(\lambda) \mathbf{L}(\lambda) \mathbf{Y}(u, v, \lambda) d\lambda, \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{3 \times H \times W}$ denotes the RGB image, $k = 1, 2, 3$ denote the red, green, and blue channels of the RGB image, (u, v, λ) denote spatial coordinate and wavelength dimension, (H, W) denote spatial dimension, respectively.

Vectorization. Let $\mathrm{vec}(\cdot)$ denote matrix vectorization, which concatenates all the columns of a matrix as a single vector. Then, $\boldsymbol{x} = \mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{3 \times n}$, where $n = H \times W$ denote spatial dimension, the original 3D spectral cube can be defined as

$$\boldsymbol{y} = \mathrm{vec}\left(\left[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{N_\lambda}\right]\right) \in \mathbb{R}^{N_\lambda \times n}. \tag{4}$$

Thus, Eq. 3 can be discretized as

$$\boldsymbol{x} = \boldsymbol{s}\boldsymbol{l}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{l} = \mathrm{vec}[\mathbf{L}] \in \mathbb{R}^{1 \times N_\lambda}$ denotes the vectorized illumination spectrum, $\boldsymbol{s} = \mathrm{vec}[\mathbf{S}] \in \mathbb{R}^{3 \times N_\lambda}$ denotes the CSS, N_λ denotes the number of sampled spectral bands, respectively. Our goal is to learn a mapping $\mathcal{G}(\cdot)$ from $\boldsymbol{x} \rightarrow \boldsymbol{y}$ with both unknown illuminations and CSSs, as

$$\hat{\boldsymbol{y}} = \mathcal{G}\left(\boldsymbol{x}\right). \tag{6}$$

Different from most existing learning-based methods that naively learn an end-to-end mapping between $\boldsymbol{x}$ and $\boldsymbol{y}$, we take both $\boldsymbol{s}$ and $\boldsymbol{l}$ into consideration, thus ensuring physical integrity.

Lin et al. (Lin & Finlayson, 2020) proposed reconstructing spectra via orthogonal subspace decomposition with known CSS while ignoring the illumination, where they demonstrate all possible solutions of $\hat{\boldsymbol{y}}$ can be decomposed into range-space components and null-space components:

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}^{\parallel} + \hat{\boldsymbol{y}}^{\perp}. \tag{7}$$

Since $\hat{\boldsymbol{y}}^{\parallel}$ is fixed that can be directly calculated from $\boldsymbol{s}$ and $\boldsymbol{x}$, thereby their goal is to estimate the null-space component $\hat{\boldsymbol{y}}^{\perp}$. Their framework shows the following limitations:

1) It requires known CSSs when calculating the range-space component $\hat{\boldsymbol{y}}^{\parallel}$, which is impractical in most real-world applications;

2) Although intensity-scaling has been applied to training data to simulate illumination variance across different samples, the real intensity of illumination depends on the exposure settings of different captures, one preset scaling factor is obviously insufficient across different samples;

3) Directly estimating the null-space component $\hat{\boldsymbol{y}}^{\perp}$ and recovering ground-truth RGBs in the same network is hard to train and may lead to suboptimal results due to trade-off between two tasks.

Remark. Note that the range-space component $\hat{\boldsymbol{y}}^{\parallel}$ plays a vital role in ensuring that the reconstruction results align with the physical degradation process. In fact, for Eq. 7, the solution $\hat{\boldsymbol{y}}^{\parallel}$ relies on a fixed projection matrix solely dependent on camera spectral sensitivities, which is clearly inadequate considering the variation of capture conditions (including capture devices, settings, and illumination conditions, and compression artefacts etc.), leading to a poorly constructed null-space $\hat{\boldsymbol{y}}^{\perp}$ that contains both physical errors and measurement/reconstruction noise, which is both difficult to model.

Orthogonal Subspace Decomposition for Spectral Reconstruction. Suppose we have the estimation of $\boldsymbol{s}$ and $\boldsymbol{l}$, then we can define a linear operator $\boldsymbol{\Phi} = \boldsymbol{s}\boldsymbol{l}$ that represents the spectral downsampling process while its pseudo-inverse $\boldsymbol{\Phi}^{\dagger}$ represents the spectral upsampling procedure. As such, Eq. 5 can be reformulated as:

$$\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{y}. \tag{8}$$

We find that orthogonal subspace decomposition is an ideal choice for ensuring the physical consistency of spectra reconstruction. By applying orthogonal subspace decomposition to Eq. 8:

$$\boldsymbol{\Phi}\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi}\boldsymbol{y} + \boldsymbol{\Phi}(\mathbf{I} - \boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi})\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{y} + \mathbf{0} = \boldsymbol{x}. \tag{9}$$

It shows that after spectral-wise downsampling, the range-space component, $\boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi}\boldsymbol{y}$, accurately reflects the RGB image $\boldsymbol{x}$. Conversely, the null-space component, $(\mathbf{I} - \boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi})\boldsymbol{y}$, does not influence the output of the downsampler $\boldsymbol{\Phi}$.

According to the observation, we can reformulate Eq. 7 into two parts: a range-space component, $\boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi}\boldsymbol{y}$, i.e., $\boldsymbol{\Phi}^{\dagger}\boldsymbol{x}$; and a null-space component, $(\mathbf{I} - \boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi})\Delta\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}^{\dagger}\boldsymbol{x} + (\mathbf{I} - \boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi})\Delta\hat{\boldsymbol{y}}, \tag{10}$$

where $\Delta\hat{\boldsymbol{y}} \in \mathbb{R}^{N_\lambda \times n}$ denotes the raw prediction of spectra signal by neural networks.

Remark. By re-parameterizing the range-space component $\hat{\boldsymbol{y}}^{\parallel}$ as a ***learnable projection*** via a new transform matrix $\boldsymbol{\Phi}$, we integrate CSS and illumination estimation as matrix coefficients. This is a much more accurate compared

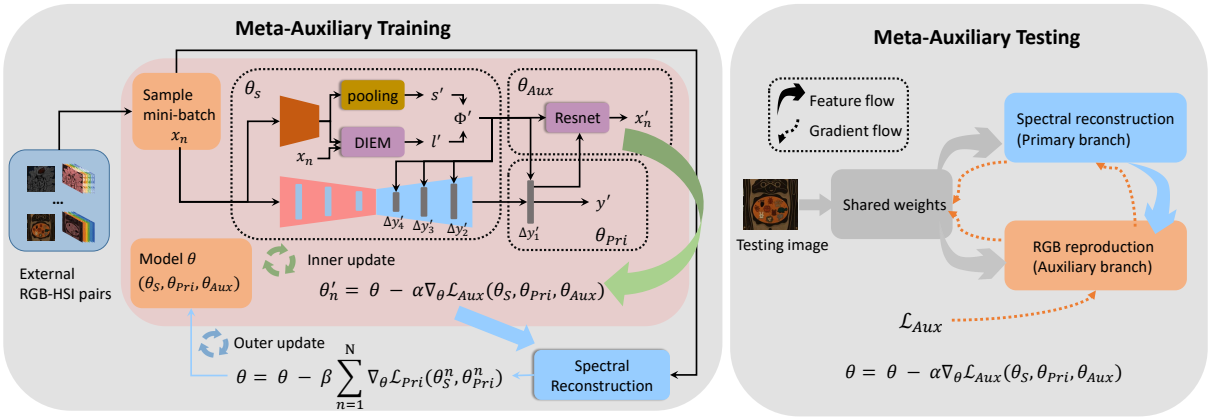

Figure 4. Overview of **PhySpec** in the meta-auxiliary learning framework: During the meta-auxiliary training phase, we first adapt the model parameters based on the auxiliary loss. The updated parameters are then evaluated on the primary task, and the model weights are ultimately refined using the primary loss derived from the adapted parameters. In the meta-testing phase, the adaptation step is applied to update the model for each test sample.

with Lin et al.'s method (Lin & Finlayson, 2020), enabling the null-space component $\hat{\boldsymbol{y}}^{\parallel}$ to focus solely on ***physically meaningful factors***, instead of having to compensate ***range-space approximation errors*** caused by different capture conditions (*e.g.*, devices, settings and illuminations): $\hat{\boldsymbol{y}}^{\perp}$ can now serve as an efficient compensation and regularization term which enhances the spectral reconstruction with nonlinear residual information that adheres to both data measurements and prior information.

3.3. Architecture

The overview of our proposed architecture is shown in Fig. 4. It contains two tasks: the primary task $\mathcal{G}(\cdot)$ takes RGB images as input to reconstruct spectra, as well as estimates the CSS and illumination descriptor; the auxiliary task $\mathcal{F}(\cdot)$ takes both estimated CSS, illumination descriptor and spectra as input to reproduce the ground-truth in a self-supervised manner. In the primary task $\mathcal{G}(\cdot)$, a UNet-based encoder-decoder (Cai et al., 2022b) architecture is utilized, and we adopt a multi-scale scheme (Yang et al., 2024) to generate spectra in different scales. In the auxiliary task $\mathcal{F}(\cdot)$, several resnet (He et al., 2016) blocks are used to reproduce the final ground-truth RGB images. Overall, most parameters of the two tasks are shared.

CSS Estimation. We utilize a transformer-based encoder (Cai et al., 2022b) as the feature extractor to extract latent feature f. A conv layer and a global average pooling layer are adapted to produce the estimated camera spectral sensitivity (CSS) $\hat{s}$.

Dynamic Illumination Estimation Module. Considering the illumination information is image-specific, using fixed parameters to learn this feature may be suboptimal, especially when adapting training parameters to unseen testing

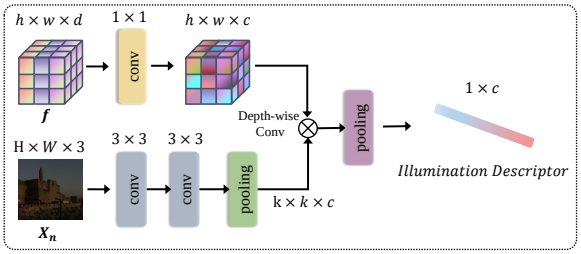

Figure 5. The architecture of the dynamic illumination estimation module. The $\boldsymbol{f}$ is the latent feature learned by the encoder, and the X_n is the input image.

samples. As such, we propose a dynamic illumination estimation module to adaptively capture the illumination-aware representation conditioned on each specific input feature. First, we employ two convolution layers to encode the input image into feature $\boldsymbol{h}$. Next, we apply average pooling to the feature $\boldsymbol{h}$ to generate an illumination-aware filter with kernel size k: $g_k(\boldsymbol{h}) \in \mathbb{R}^{k \times k \times c}$. At last, the latent feature $\boldsymbol{f}$ is convolved with the illumination-aware filter with depth-wise convolution to obtain an illumination-aware representation:

$$\boldsymbol{l}' = \boldsymbol{f} \otimes g_k(\boldsymbol{f}). \tag{11}$$

Note that the illumination descriptor $\boldsymbol{l}'$ is implicitly estimated since there is no illumination information provided by the training dataset. Besides, it is also impractical to access such ground-truth illumination information in the real world.

Objective Function. As illustrated in Fig. 4, the network parameters θ are divided into three components: θ_S, θ_{Pri}, and θ_{Aux}. In this structure, θ_S denotes the shared parameters, while θ_{Pri} and θ_{Aux} correspond to the parameters specific to the primary and auxiliary tasks. The output from

Algorithm 1 Meta-auxiliary Training

Input: $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s})$ triples, λ, α, β: learning rates
Output: θ: meta-auxiliary learned parameters
Pre-train the whole model: $\theta = \{\theta_S, \theta_{Pri}, \theta_{Aux}\}$
while not converged **do**
  Sample whole training data $\{\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{s}^k\}_{k=1}^K$
  Evaluate pre-training loss $\mathcal{L}_{Pre}$ by Eqn. 14
  Update θ: $\tilde{\theta} \leftarrow \theta - \lambda \nabla_\theta \mathcal{L}_{Pre}\left(\theta_S, \theta_{Pri}, \theta_{Aux}\right)$
end
Initialize the model with the above pre-trained weights
while not converged **do**
  Sample a mini-batch of training triple $\{\boldsymbol{x}^n, \boldsymbol{y}^n, \boldsymbol{s}^n\}_{n=1}^N$
  for each n **do**
    Evaluate auxiliary loss $\mathcal{L}_{Aux}$ by Eqn. 13
    Compute adapted parameters θ^n with gradient descent: $\tilde{\theta}_n \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{Aux}\left(\theta_S, \theta_{Pri}, \theta_{Aux}\right)$
    Update: $\theta_{Aux} \leftarrow \theta_{Aux} - \alpha \nabla_\theta \mathcal{L}_{Aux}\left(\theta_{Aux}\right)$
  end
  Validate the primary task and update:
  $\theta \leftarrow \theta - \beta \sum_{n=1}^N \nabla_\theta \mathcal{L}_{Pri}^n\left(\theta_S^n, \theta_{Pri}^n\right)$
end

the final shared layer is directed into two branches; one is responsible for generating the final spectra $\hat{\boldsymbol{y}}$ (the primary task), and the other uses $\hat{\boldsymbol{y}}$ as an additional input to reconstruct the original RGB images. This approach allows that during the testing phase, the parameters of the primary task are updated exclusively based on the auxiliary loss. For both tasks, we employ the mean relative absolute error (MRAE) as the loss function:

$$\mathcal{L}_{Pri}\left(\theta_S, \theta_{Pri}\right) = \|\boldsymbol{s} - \hat{\boldsymbol{s}}\|_1 + \sum_{i=1}^4 \left\|\boldsymbol{y}^i - \hat{\boldsymbol{y}}^i\right\|_1, \quad (12)$$

$$\mathcal{L}_{Aux}\left(\theta_S, \theta_{Pri}, \theta_{Aux}\right) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1. \quad (13)$$

Then, we pre-train the whole network on the training dataset, using the combination loss as:

$$\mathcal{L}_{Pre}(\theta) = \mathcal{L}_{Pri}\left(\theta_S, \theta_{Pri}\right) + \mathcal{L}_{Aux}\left(\theta_S, \theta_{Pri}, \theta_{Aux}\right). \quad (14)$$

3.4. Meta-Auxiliary Learning

MAXL aims to integrate both external and internal learning to enable rapid adaptation of trained parameters to unseen samples, requiring only a limited number of gradient descent steps during testing stage. In our scenario, we further introduce MAXL to enforce generated HSIs to accurately recover ground-truth RGBs in a self-supervised manner.

Meta-Auxiliary Training. Given a triple of training images $(\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{s}_n)$ and the pre-trained model θ, we first adapt θ using the auxiliary loss with only several gradient descent updates

$$\tilde{\theta}_n \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{Aux}\left(\theta_S, \theta_{Pri}, \theta_{Aux}\right), \quad (15)$$

where α denotes the adaptation learning rate. Note that the adaptation step involves all parameters (*i.e.*, θ_S, θ_{Pri}, and θ_{Aux}) with only $\boldsymbol{x}_n$ utilized.

To effectively adapt the pre-trained parameters θ for testing, it is essential to update θ_S and θ_{Pri} corresponding to the primary task. Consequently, we can define the meta-objective as follows:

$$\arg\min_{\theta_S, \theta_{Pri}} \sum_{n=1}^N \mathcal{L}_{Pri}^k\left(\theta_S^k, \theta_{Pri}^k\right), \quad (16)$$

where N is the number of training samples. Note that the meta-objective in Eq. 16 can be minimized via gradient descent

$$\theta \leftarrow \theta - \beta \sum_{n=1}^N \nabla_\theta \mathcal{L}_{Pri}^k\left(\theta_S^k, \theta_{Pri}^k\right), \quad (17)$$

where β denotes the meta-learning rate. In practice, we use a mini-batch instead of the entire external training dataset to update Eq. 17. The algorithm is given in Alg. 1. Note that θ_S and θ_{Pri} are updated in the outer loop, whereas θ_{Aux} is updated in the inner loop.

Meta-Auxiliary Testing. At the testing phase, we simply fine-tune the meta-learned parameters using merely several steps of gradient descent on a testing sample $\boldsymbol{x}$ with Eq. 15, as illustrated in Alg. 2.

Algorithm 2 Meta-auxiliary Testing

Input: $\boldsymbol{x}$: a testing sample
  k: gradient updating steps
  α: learning rate of testing adaptation
Output: Reconstructed hyperspectral image $\hat{\boldsymbol{y}}$
Initialize model parameter θ with pre-trained weights
for k steps **do**
  Evaluate auxiliary loss $\mathcal{L}_{Aux}$ by Eqn. 13
  Update $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{Aux}(\theta_S, \theta_{Pri}, \theta_{Aux})$
end
return $\hat{\boldsymbol{y}}$ from Eqn. 7

4. Experiments

4.1. Datasets and Implementation Details

Datasets. To evaluate the generalization and the fidelity of our method, we conduct experiments on three HSI reconstruction datasets (*e.g.*, the ARAD-1K Synthetic dataset, the ARAD-1K Real dataset (Arad et al., 2022) and the ICVL HSI dataset (Arad & Ben-Shahar, 2016)). The ARAD-1K

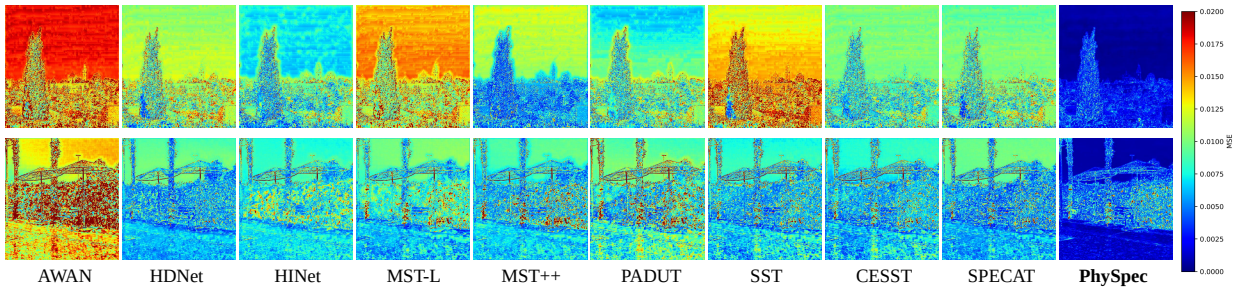

Figure 6. The MSE error map obtained from the validation subset of the ARAD-1K Real dataset, which is calculated along the spectral direction, showcasing the discrepancies between the reconstructed HSIs and the corresponding ground truths.

Table 1. Quantitative evaluations. All compared methods are trained on the ARAD-1k Synthetic dataset and ICVL dataset, while directly evaluated on the ARAD-1K Real data.

Method	Params (M)	FLOPs (G)	ARAD-1K Synthetic			ARAD-1K Real			ICVL		
			SAM↓	SSIM↑	PSNR↑	SAM↓	SSIM↑	PSNR↑	SAM↓	SSIM↑	PSNR↑
AWAN*(CVPRW'20) (Li et al., 2020)*	4.04	270.61	8.05	0.917	33.15	16.72	0.896	30.36	4.59	0.918	31.92
HDNet*(CVPR'22) (Hu et al., 2022)*	2.66	173.81	7.31	0.922	33.54	14.40	0.899	31.08	4.17	0.924	31.85
HINet*(CVPR'21) (Chen et al., 2021)*	5.21	31.04	7.04	0.930	33.91	12.11	0.903	31.35	4.19	0.928	32.01
MST-L*(CVPR'2022) (Cai et al., 2022a)*	2.45	32.07	5.17	0.935	34.68	10.36	0.906	31.87	3.51	0.935	32.87
MST++*(CVPRW'22) (Cai et al., 2022b)*	1.62	23.05	4.87	0.939	34.93	9.62	0.910	32.05	3.04	0.941	32.44
PADUT*(ICCV'23) (Li et al., 2023b)*	6.38	107.15	4.34	0.952	35.49	9.03	0.912	31.97	3.11	0.948	33.07
SST*(ArXiv'23) (Cai et al., 2024)*	12.74	219.38	4.82	0.941	35.05	8.31	0.915	31.60	3.71	0.935	32.71
CESST*(AAAI'24) (Yang et al., 2024)*	1.54	90.18	4.97	0.945	36.05	7.19	0.923	32.15	3.27	0.939	32.96
SPECAT*(CVPR'24) (Yao et al., 2024)*	**0.37**	15.97	4.62	0.940	35.45	8.04	0.918	31.74	3.81	0.944	32.54
PhySpec***(ours)***	2.74	**4.65**	**4.12**	**0.967**	**37.12**	**4.17**	**0.937**	**33.87**	**1.95**	**0.957**	**35.04**

dataset, the largest in this domain, comprises 950 RGB-HSI pairs (482 × 512 resolution, 31 spectral channels from 400nm to 700nm), with 900 used for training and 50 for validation. The ARAD-1K Synthetic dataset is constructed by randomly selecting 23 of Jiang et al.'s CSSs (Jiang et al., 2013) for training input generation and reserving 5 for testing. The ARAD-1K Real dataset (Arad et al., 2022) retains its original setting with unknown CSSs and compression patterns. The ICVL dataset consists of 201 high-resolution HSIs. Since paired RGB images are not provided, we generate them following Magnusson et al.'s method (Magnusson et al., 2020). To ensure consistency, we exclude 18 images with varying resolutions, utilizing 147 image pairs for training and 36 for testing.

Implementation Details. The proposed PhySpec is implemented with Pytorch. All images are linearly rescaled within the range between 0 and 1. During training, images are cropped into 128×128 pixel patches with a stride of 8. Data augmentation is applied through random flipping and rotation. The batch size is fixed at 20. During the pre-training stage, we utilize the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 10^{-4}, following a Cosine Annealing schedule (Loshchilov & Hutter, 2016) over 300 epochs to facilitate gradual learning rate decay. During the meta-auxiliary learning phase, we set the parameters α and β to 1×10^{-2} and 5×10^{-5}, respectively, to balance the learning dynamics between the primary and auxiliary tasks. For test-time adaptation, four gradient descent updates are performed. All experiments are conducted on a single NVIDIA Ampere A100 with 40G RAM.

4.2. Quantitative Results

We evaluate our method against nine state-of-the-art (SOTA) approaches, categorized as follows: three RGB-based reconstruction methods (AWAN (Li et al., 2020), MST++ (Cai et al., 2022b) and CESST (Yang et al., 2024)); five CASSI-based reconstruction methods (HDNet (Hu et al., 2022), MST-L (Cai et al., 2022a), PADUT (Li et al., 2023b), SST (Cai et al., 2024), SPECAT (Yao et al., 2024)); and one image restoration method (HINet (Chen et al., 2021)). We evaluate the performance using three widely-used metrics, including SAM, SSIM, and PSNR. The first one assesses spectral quality, while the latter two evaluate spatial quality. Lower SAM values indicate better spectral quality, while higher SSIM and PSNR values signify better spatial quality. As shown in Table 1, our method achieves the best performance overall metrics on both the ARAD-1K synthetic dataset and ARAD-1K real dataset, as well as the ICVL dataset. To evaluate the model complexity, we also compare parameters (spatial complexity) and FLOPs (temporal complexity) in Table 1. As can be seen, our method achieves the lowest FLOPs cost with relatively low parameters.

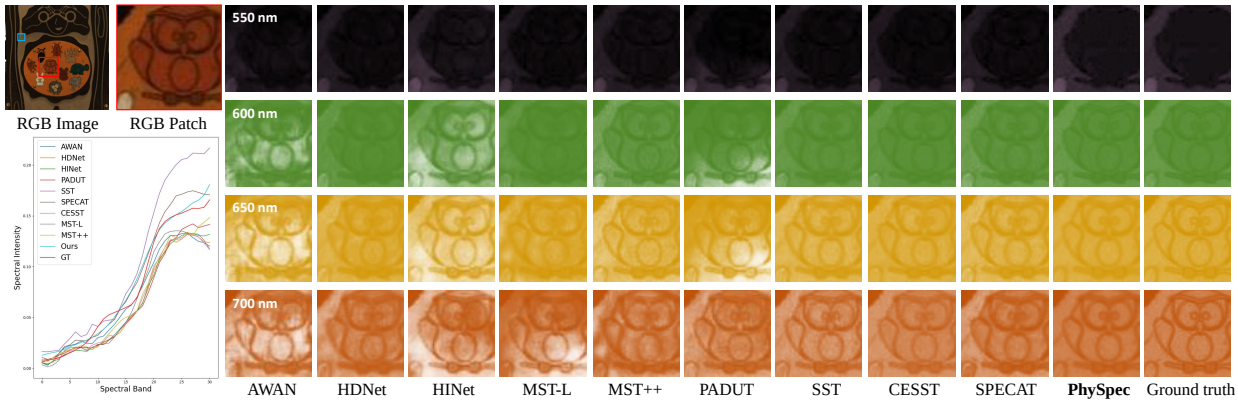

Figure 7. Visual comparisons on a randomly selected scene from the validation set of the ARAD-1K Real dataset with 4 spectral channels. The bottom-left spectral curves represent the region highlighted in blue in the RGB image. For better visual comparison, please zoom in.

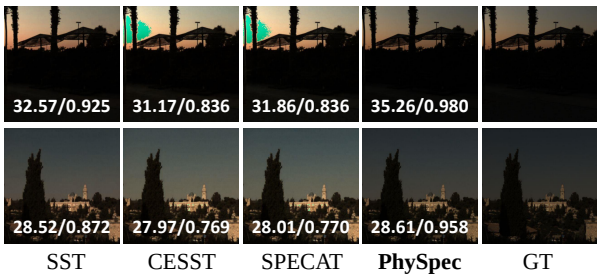

Figure 8. Reproduced RGB images from reconstructed HSIs with PSNR and SSIM metrics.

4.3. Qualitative Results

Visual comparisons are provided in Fig. 6 and Fig. 7. Fig. 6 illustrates spectral-dimension MSE error maps between generated and ground-truth HSIs. Fig. 7 displays reconstructed hyperspectral images across selected spectral channels (550nm, 600nm, 650nm, and 700nm). It is obvious that the existing HSI reconstruction methods exhibit limitations in estimating spectral intensity and recovering fine details, particularly in high-frequency regions (e.g., sky textures). In contrast, PhySpec achieves enhanced spatial detail restoration and pixel-level smoothness. Additionally, the spectral density curves, shown in the bottom-left of Fig. 7, correspond to the selected region (blue box) in the RGB image (top-left). The strong correlation between our curve and the ground truth shows PhySpec's ability to achieve spectrally consistent reconstruction.

4.4. Ablation Studies

Is PhySpec really physically consistent? Our PhySpec formulates a self-reconstruction framework to ensure the physical consistency, which is inspired by CycleGAN (Zhu et al., 2017) and CoColor (Yang et al., 2023). To validate the physically consistent property of PhySpec, we provide the reproduced RGB comparison in Fig. 8 with PSNR and SSIM metrics. As can be seen, CESST and SPECAT fail to predict artifact-free results, while SST fails to reproduce consistent tune and brightness with ground truths. In contrast, our method reproduces results that are consistent with ground truths and achieves high quantitative results of both PSNR and SSIM, indicating the physical consistency of our model.

Table 2. The top sub-table illustrates the break-down ablations of our proposed PhySpec. The bottom sub-table investigates the gradient descent steps in the meta-auxiliary testing.

Method	DIEM	CSS	MAXL Training	MAXL Testing	PSNR	SAM
Sim-PhySpe					34.67	5.12
Variant 1	✓				34.85	4.98
Variant 2	✓	✓			36.10	4.71
Variant 3	✓	✓	✓		36.95	4.20
PhySpec	✓	✓	✓	✓	**37.12**	**4.12**

Setting	k	PSNR	SAM
PhySpec	0	36.95	4.20
PhySpec	1	36.98	4.17
PhySpec	2	37.04	4.15
PhySpec	3	37.04	4.14
PhySpec	4	**37.12**	**4.12**
PhySpec	5	**37.12**	4.14
PhySpec	6	37.11	4.13

Is PhySpec really effective? To evaluate the effectiveness of PhySpec, we conducted a breakdown of the ablation studies for the DIEM, CSS explicit estimation mechanism, meta-auxiliary training and testing adaptation scheme to analyze the impact of each part. As shown in the top sub-table in Table 2, the Sim-PhySpec used as a baseline for comparison is the pure MST structure proposed by (Cai et al., 2022a). Variant 1 disables the CSS explicit estimation mechanism by implicitly estimating CSS without ground-truth CSS supervision. The results indicate that the CSS explicit estimation and meta-auxiliary training contribute the most to performance improvement, yielding a 1.25 dB and 0.85 dB increase in PSNR, respectively. This highlights the

effectiveness of the integration of both physical constraints and self-supervised meta-auxiliary learning.

Additionally, we also investigate the effect of gradient descent update step k as reported in the bottom sub-table of Table 2. We We also examined the impact of the gradient descent update step, denoted as k, as indicated in the bottom sub-table of Table 2. We determined that $k = 4$ provides the best performance. Increasing the number of update steps could result in overfitting on the auxiliary task.

Can PhySpec really explicitly estimate CSSs? To investigate the effectiveness of CSS estimation, we provide the generated CSSs from five selected testing cameras with corresponding ground truths in Fig. 2. As can be seen, PhySpec can accurately estimate CSSs that are unseen during training, such as Nikon D700, Nikon D40, and Pentax Q, reinforcing the reliability and robustness of our method, and highlighting its potential for real-world applications.

5. Conclusion

In this paper, we introduced PhySpec, a novel method addressing the colorimetric dilemma in HSI reconstruction—a critical challenge where existing approaches fail to maintain physical consistency between reconstructed spectra and ground-truth RGB colors. PhySpec embeds imaging physics into learning via two innovations: (1) orthogonal subspace decomposition to model intrinsic HSI-RGB relationships, enabling precise CSS estimation and physically consistent spectral recovery; and (2) a self-supervised MAXL strategy that adapts model parameters to unseen test data, ensuring robust generalization while enforcing RGB fidelity as a physical constraint. Experiments demonstrate PhySpec's superior performance over state-of-the-art methods, resolving RGB-HSI mismatches and enhancing practical reliability. By bridging data-driven learning with physical principles, our work establishes a foundation for trustworthy HSI reconstruction in applications like remote sensing and computational photography. Future research will extend this framework to dynamic imaging and multi-sensor fusion.

Impact Statement

This work has potential applications in remote sensing, medical imaging, and computational photography, where accurate spectral data is essential. Ensuring physical consistency mitigates risks of inaccurate predictions for practical imaging applications. By integrating physical integrity into machine learning, PhySpec advances high-fidelity AI solutions in imaging. Future research could extend this paradigm to dynamic imaging and multi-sensor fusion, expanding its impact across diverse practical domains.

References

Arad, B. and Ben-Shahar, O. Sparse recovery of hyperspectral signal from natural rgb images. In *European conference on computer vision*, 2016.

Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., et al. Ntire 2022 spectral recovery challenge and data set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 863–881, 2022.

Borengasser, M., Hungate, W. S., and Watkins, R. *Hyperspectral remote sensing: principles and applications*. CRC press, 2007.

Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., and Van Gool, L. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17502–17511, 2022a.

Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., and Van Gool, L. MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 745–755, 2022b.

Cai, Z., Liu, Z., Yu, J., Zhang, Z., Da, F., and Jin, C. Reversible-prior-based spectral-spatial transformer for efficient hyperspectral image reconstruction. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 20(1):1–22, 2024.

Chen, L., Lu, X., Zhang, J., Chu, X., and Chen, C. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 182–192, 2021.

Cheng, Y., Wang, X., Ma, Y., Mei, X., Wu, M., and Ma, J. General hyperspectral image super-resolution via meta-transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Chi, Z., Wang, Y., Yu, Y., and Tang, J. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9137–9146, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Hu, X., Cai, Y., Lin, J., Wang, H., Yuan, X., Zhang, Y., Timofte, R., and Van Gool, L. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17542–17551, 2022.

Huang, S., Ren, M., Yang, Y., Wang, X., and Wei, Y. Mftn: A multi-scale feature transfer network based on IMatchFormer for hyperspectral image super-resolution. In *Forty-first International Conference on Machine Learning*, 2024.

Huo, D., Wang, J., Qian, Y., and Yang, Y.-H. Learning to recover spectral reflectance from RGB images. *IEEE Transactions on Image Processing*, 2024.

Jiang, J., Liu, D., Gu, J., and Süsstrunk, S. What is the space of spectral sensitivity functions for digital color cameras? In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 168–179, 2013.

Johnson, W. R., Wilson, D. W., Fink, W., Humayun, M. S., and Bearman, G. H. Snapshot hyperspectral imaging in ophthalmology. *Journal of biomedical optics*, 12(1): 014036, 2007.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klema, V. and Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980.

Lam, A. and Sato, I. Spectral modeling and relighting of reflective-fluorescent scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1452–1459, 2013.

Li, J., Wu, C., Song, R., Li, Y., and Liu, F. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 462–463, 2020.

Li, J., Leng, Y., Song, R., Liu, W., Li, Y., and Du, Q. Mformer: Taming masked transformer for unsupervised spectral reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023a.

Li, M., Fu, Y., Liu, J., and Zhang, Y. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12959–12968, 2023b.

Lin, Y.-T. and Finlayson, G. D. Physically plausible spectral reconstruction from rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 532–533, 2020.

Liu, S., Long, M., Wang, J., and Jordan, M. I. Generalized zero-shot learning with deep calibration network. *Advances in Neural Information Processing Systems*, 31, 2018.

Liu, S., Davison, A., and Johns, E. Self-supervised generalisation with meta auxiliary learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Lu, G. and Fei, B. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901, 2014.

Magnusson, M., Sigurdsson, J., Armansson, S. E., Ulfarsson, M. O., Deborah, H., and Sveinsson, J. R. Creating rgb images from hyperspectral images using a color matching function. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2045–2048. IEEE, 2020.

Robles-Kelly, A. Single image spectral reconstruction for multimedia applications. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 251–260, 2015.

Shi, Z., Chen, C., Xiong, Z., Liu, D., and Wu, F. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 939–947, 2018.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems*, 26, 2013.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248, 2020.

Wang, Y., Hu, Y., Yu, J., and Zhang, J. GAN prior based null-space learning for consistent super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2724–2732, 2023.

Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., and Wu, F. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In *ICCVW*, pp. 518–525, 2017.

Yang, X., Chen, J., and Yang, Z. Cooperative colorization: Exploring latent cross-domain priors for nir image spectrum translation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2409–2417, 2023.

Yang, X., Chen, J., and Yang, Z. Hyperspectral image reconstruction via combinatorial embedding of cross-channel spatio-spectral clues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6567–6575, 2024.

Yang, X., Chen, J., and Yang, Z. Learning physics-informed color-aware transforms for low-light image enhancement. *arXiv preprint arXiv:2504.11896*, 2025.

Yao, Z., Liu, S., Yuan, X., and Fang, L. Specat: Spatial-spectral cumulative-attention transformer for high-resolution hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25368–25377, 2024.

Yuan, Y., Zheng, X., and Lu, X. Hyperspectral image superresolution by transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):1963–1974, 2017.

Zhang, L., Nie, J., Wei, W., and Zhang, Y. Unsupervised testtime adaptation learning for effective hyperspectral image super-resolution with unknown degeneration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5008–5025, 2024.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Zhu, Z., Liu, H., Hou, J., Zeng, H., and Zhang, Q. Semantic-embedded unsupervised spectral reconstruction from single rgb images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2279–2288, 2021.