ColorMamba: Towards High-quality NIR-to-RGB Spectral Translation with Mamba

Huiyu Zhai[1] WENYU.ZHY@GMAIL.COM
Guang Jin[1] JINGUANG720407@GMAIL.COM
Xingxing Yang[2*] CSXXYANG@COMP.HKBU.EDU.HK
Guosheng Kang[1] GUOSHENGKANG@GMAIL.COM
[1] *School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China*
[2] *Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR*

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Translating NIR to the visible spectrum is challenging due to cross-domain complexities. Current models struggle to balance a broad receptive field with computational efficiency, limiting practical use. Although the Selective Structured State Space Model, especially the improved version, Mamba, excels in generative tasks by capturing long-range dependencies with linear complexity, its default approach of converting 2D images into 1D sequences neglects local context. In this work, we propose a simple but effective backbone, dubbed ColorMamba, which first introduces Mamba into spectral translation tasks. To explore global long-range dependencies and local context for efficient spectral translation, we introduce learnable padding tokens to enhance the distinction of image boundaries and prevent potential confusion within the sequence model. Furthermore, local convolutional enhancement and agent attention are designed to improve the vanilla Mamba. Moreover, we exploit the HSV color to provide multi-scale guidance in the reconstruction process for more accurate spectral translation. Extensive experiments show that our ColorMamba achieves a 1.02 improvement in terms of PSNR compared with the state-of-the-art method. Our code is available at https://github.com/AlexYangxx/ColorMamba/.

Keywords: NIR Image; Spectral Translation; Colorization; Mamba; State Space Model.

1. Introduction

"*What I cannot create, I do not understand.*"

—— Richard P. Feynman, 1988

For many years, the pursuit of transcending mere observation to achieve a profound comprehension of visual data has driven souls to the art of generation itself. Early attempts, such as variational autoencoders (Kingma and Welling, 2013) and generative adversarial networks (Goodfellow et al., 2020), have shown impressive performance in various downstream tasks, for example, image super-resolution (Saharia et al., 2022) and grayscale image colorization (Su et al., 2020). In this work, we focus on studying a niche but important downstream generative task, NIR-to-RGB spectral translation.

The Near-Infrared (NIR) spectrum (780nm – 1000nm) is adjacent to the visible spectrum (380nm – 780nm), which is invisible to human eyes. Owing to the longer wavelength

* corresponding author

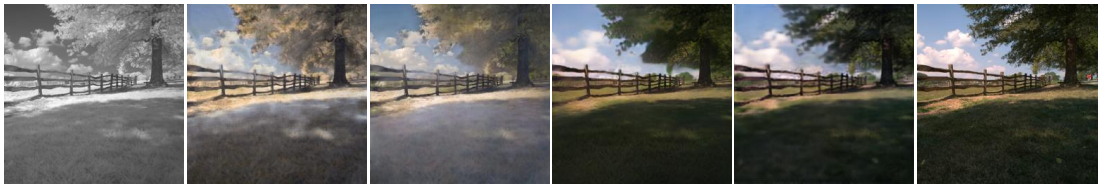

Figure 1: **Visual effect display** compared to three methods: Restorer (Zamir et al., 2022), DRSformer (Chen et al., 2023a), and CoColor (Yang et al., 2023a).

compared with traditional RGB imaging, NIR imaging has been widely applied in object detection (Takumi et al., 2017), nighttime video surveillance (Christnacher et al., 2018), remote sensing (Thenkabail et al., 2018) and so on. Nevertheless, the spectral response encapsulated within NIR imagery diverges substantially from the perceptual experiences familiar to both human observers and computer vision systems, both of which have been predominantly attuned to scene reflectance within the visible-light spectrum. To render the visualization of NIR images more congruent with innate human perception and intuitive interpretation, the field of NIR-to-RGB spectral domain translation has emerged as a subject of considerable academic interest.

Existing NIR-to-RGB spectral translation methods mainly focus on learning the pixel-wise mapping relations and utilize U-Net as the backbone for this dense prediction task (Yang and Chen, 2020; Yang et al., 2023b,a; Zhai et al., 2024). However, these methods are all based on CNN structures, whose receptive fields are local. Yet, the marked improvements seen in recent advanced deep learning models for image translation (Yang et al., 2024; Guo et al., 2024) tasks have been significantly ascribed to the enlarged receptive fields within the neural networks, allowing for broader contextual understanding in processing images. Subsequently, translation methods based on Transformer architectures (Dosovitskiy et al., 2020), which inherently exhibit global receptive fields, have demonstrably surpassed the performance of those based on Convolutional Neural Network (CNN) frameworks in empirical assessments. Corroborated by this observation, recent literature (Chen et al., 2023b; Guo et al., 2024) suggests a positive correlation between the extent of pixel activation within such models and the resulting improvements in translation outcomes.

Recent advances in the domain of structured state-space sequence models (S4) (Gu et al., 2021a), and more specifically their refined variant known as Mamba (Gu and Dao, 2023), have been recognized for their efficiency and efficacy as foundational models for the development of deep neural networks (Sze et al., 2017). Such progress holds promise for reconciling the expansive global receptive fields demanded for high-quality image generation with the imperative of computational efficiency. In addition, the implementation of the parallel scan algorithm (Sengupta et al., 2011) permits Mamba-based models to process discrete elements, or tokens, concurrently, thereby optimizing the utilization of contemporaneous computational apparatuses such as GPUs. The aforementioned advantageous characteristics compel further exploration of Mamba's capacity for facilitating efficient and far-reaching modeling within architectures engineered for image generation.

Nevertheless, the vanilla Mamba model (Gu and Dao, 2023), does not seamlessly adapt to the challenges presented by spectral translation contexts. Principally, it processes images as flattened one-dimensional sequences via recursive computation, which can inadvertently dislocate spatially proximate pixels to disparate positions within the one-dimensional array, thereby giving rise to a phenomenon identified as ***local context neglect***, where the spatial correlation between adjacent pixels is not adequately preserved.

To address these challenges, we proposed ColorMamba, which is the first work to introduce Mamba into spectral translation tasks. Specifically, considering that directly applying vanilla Mamba to 2D vision tasks will result in local context neglect (Guo et al., 2024), we introduce local convolutional enhancement and agent attention (Han et al., 2023) mechanisms to evolve as Visual State Space Blocks (VSSBs), which can model both long-range dependencies and local contexts. Moreover, to enhance the distinction of image boundaries and prevent potential confusion within the sequence model, a new scanning strategy is proposed. Specifically, we insert learnable padding tokens between two adjacent tokens in the scanning sequence of the state space model that do not share a proximate spatial correlation. This strategically inserted padding facilitates the Mamba blocks in more accurately interpreting image peripheries, thereby reinforcing the model's spatial awareness and the integrity of the sequential data processing. Besides, we further propose an Hue-Saturation-Value (HSV) color prediction sub-network to exploit HSV color prior, which serves as multi-scale guidance in the reconstruction process of RGB predictions.

In summary, our contributions lie in the following aspects:

- We propose a Mamba-based backbone for NIR-to-RGB spectral translation, dubbed ColorMamba, capable of modeling both global long-range dependencies and capturing local contexts. To our best knowledge, our ColorMamba is the first Mamba-based method for spectral translation.

- We evolve the vanilla Mamba with local convolutional enhancement and agent attention, as well as a new scan strategy, to formulate our Visual State Space Blocks (VSSBs), which address the local context neglect dilemma efficiently and boost the performance of standard Mamba on 2D images.

- We propose an HSV color prediction sub-network that exploits color prior to provide multi-scale guidance in the reconstruction process for more accurate spectral translation.

- Extensive experimental results show that our ColorMamba achieves a 1.02 improvement in terms of PSNR compared with the state-of-the-art method, which suggests that our ColorMamba offers a potent and auspicious foundational architecture for endeavors in spectral translation.

2. Related Work

2.1. Spectral Translation

Notwithstanding a degree of congruence in the colorization processes of grayscale and Near-Infrared (NIR) imagery, the task of NIR-to-RGB spectral translation presents considerably

greater challenges due to the significant spectral domain disparities and the paucity of labeled data. Further complexity is introduced by environmental variations, such as thermal changes and alterations in the light source, which can lead to substantial variances in the intensity of NIR images captured even within identical settings, thereby exacerbating the ambiguity in mapping. Initial methodologies have typically employed Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) to transform NIR images to the color space, relying on pixel-level supervision from ground-truth RGB images (Suárez et al., 2018, 2017). However, such approaches frequently resulted in color distortions and blurred outputs, with the learning process being constrained by the scarcity of NIR-RGB image pairs and local receptive fields. In response, unsupervised and semi-supervised strategies have been proposed to capitalize on unpaired data (Mehri and Sappa, 2019; Yang and Chen, 2020), predominantly employing CycleGAN frameworks (Zhu et al., 2017) to translate RGB images into plausible NIR counterparts, thus augmenting the limited training dataset. Regrettably, the '*RGB-to-NIR-to-RGB*' transition process inevitably resulted in the loss of rich textural information inherent within the NIR domain, owing to the spectral discrepancy present between NIR and the visible spectrum. Very recently, Yang et al. (Yang et al., 2023b) proposed a multi-scale progressive learning framework that utilized domain adaptation techniques to incorporate grayscale image colorization to disambiguate the mapping relations between NIR and RGB domains. Following this work, Yang et al. proposed a cooperative learning (Yang et al., 2023a) paradigm that leveraged grayscale prior knowledge to guide the colorization of NIR images. However, incorporating grayscale images into spectral translation introduces a multi-task learning problem that is difficult to train.

2.2. State Space Models

State Space Models (SSMs) (Gu et al., 2021b,a) is a mathematical model of a physical system specified as a set of inputs, outputs, and variables. Very recently, SSMs have been introduced into the domain of deep learning, emerging as a powerful substitute to CNN- and Transformer-based backbones, benefiting from their notable property of linear complexity concerning sequence length in the modeling of long-range dependencies. Early efforts were marked by the inception of the Structured State-Space Sequence model (S4) (Gu et al., 2021a), which pioneered the deep state-space approach to long-range dependency modeling.

Recently, Mamba (Gu and Dao, 2023), a data-dependent SSM characterized by a selective mechanism and an optimally designed hardware framework, has outshined Transformer models in natural language processing tasks, maintaining linear complexity relative to input length. Before Mamba's advent, the application of SSMs in computer vision was exemplified by DiffuSSM (Yan et al., 2024), the first diffusion model that uses SSMs as a substitute for attention mechanisms. Mamba has since raised the bar for modeling efficiency compared to its predecessors, inspiring an array of Mamba variants geared towards image- and video-based vision tasks (Islam et al., 2023; Wang et al., 2023; Liu et al., 2024). Currently, innovative approaches have incorporated Mamba into image generation, with works such as DiS (Fei et al., 2024) exploring its generative potential at resolutions up to 512×512 by integrating the ViM variant (Zhu et al., 2024). Additionally, ZigMa (Hu et al., 2024) leverages vanilla Mamba blocks alongside diverse scan patterns, training on high-resolution human facial generation datasets (Karras et al., 2019), illustrating the versatility and applicability

of the Mamba architecture across various domains within image processing and generation. In the current study, we extend the potential applications of the Mamba model to the field of spectral translation. Our investigation proposes a straightforward yet efficacious benchmark, setting a foundation for future research endeavors in this area.

3. Methodology

3.1. Preliminaries

Vision State Space Module can exploit long-range dependencies with global receptive fields by processing 2D images as flattened 1D sequences via recursive computation. However, the vanilla scanning strategy utilized by the standard Mamba will inadvertently dislocate spatially proximate pixels to disparate positions within the one-dimensional array, thereby giving rise to a phenomenon identified as local context neglect. We propose learnable padding tokens between two adjacent tokens in the scanning sequence of the state space model that do not share a proximate spatial correlation to enhance the distinction of image boundaries and prevent potential confusion within the sequence model. Figure 2 provides a detailed illustration of the method. Specifically, given the input feature map $\mathbf{F} \in \mathbb{R}^{1\times H\times W\times N}$, we first pad the feature map into $\mathbf{F_p} \in \mathbb{R}^{1\times (H+2)\times (W+2)\times N}$. Then, we unfold the feature map by transforming the two-dimensional spatial information into a set of four one-dimensional sequences, each containing $(H+2)(W+2)$ elements. The reorganization process engages four distinct scanning paths, including diagonal orientations: top left to bottom right (left-right), top left to bottom right (top-down), bottom right to top left (right-left), and bottom right to top left (down-top) to capture the spatial continuum of the feature map effectively. These restructured sequences are formally denoted as $\{\mathbf{S}_d \in \mathbb{R}^{1\times L\times N}\}_{d=1}^{n}$, where $n = 4$ is the count of sequences and $L = (H+2)(W+2)$ denotes the length of each sequence.

According to the continuous linear time-invariant systems, we can map a 1D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ via an implicit latent state $h(t) \in \mathbb{R}^N$, which can be rigorously formulated as a linear ordinary differential equation (ODE):

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t), \end{aligned} \tag{1}$$

where N is the state size, $h'(t)$ is the derivative of h, $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, and $\mathbf{D} \in \mathbb{R}$, are the weights. Typically, a discretization process is needed to apply the SSM to 2D visual signals. Especially, Mamba leverages the zero-order hold (ZOH) rule to discretize Eq. 1 as:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \end{aligned} \tag{2}$$

where Δ denotes the timescale parameter. Now the discretized version of Eq. 1 based on restructured sequences $\{\mathbf{S}_d \in \mathbb{R}^{1\times L\times N}\}_{d=1}^{n}$ can be formulated in a recursive form:

$$\begin{aligned} h_k^d &= \overline{\mathbf{A}}h_{k-1}^d + \overline{\mathbf{B}}\mathbf{S}_d, \\ y_k^d &= \mathbf{C}h_k^d + \mathbf{D}\mathbf{S}_d. \end{aligned} \tag{3}$$

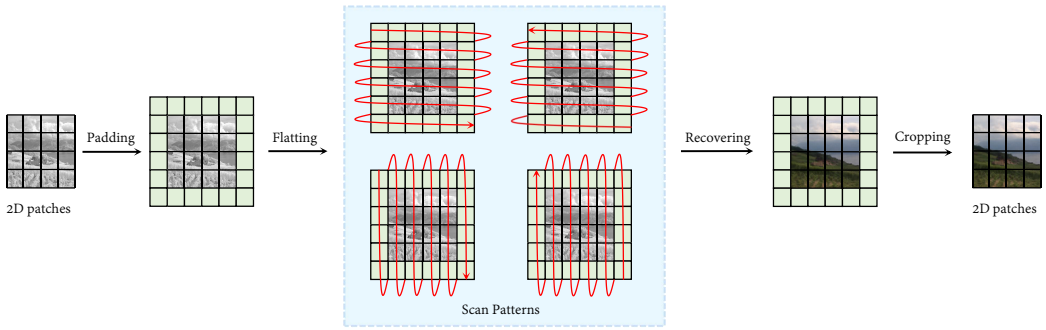

Figure 2: **Status scanning strategy.** We inject learnable padding tokens between two adjacent tokens that do not share approximate spatial correlation to enhance boundary distinction and prevent potential confusion within the sequence model.

Then, we merge all sequence features $\{\mathbf{y}_k^d\}_{d=1}^n$ to get the output map $\mathbf{y} = \sum_{d=1}^n \mathbf{y}_k^d$. Finally, the output map is cropped into the original dimension before padding.

3.2. Overall Architecture

In this study, we introduce a novel backbone architecture designed for spectral translation, which processes a monochromatic NIR image ($x_{nir} \in \mathbb{R}^{H\times W\times 1}$) as input and yields a colorized NIR image ($y_{rgb} \in \mathbb{R}^{H\times W\times 3}$) as output. Acknowledging the importance of transferring color details from RGB ground truths to NIR inputs, we integrate an HSV Color Prediction Sub-network (denoted as G_B), intended to provide a robust and dynamic color prior, thereby assisting the primary RGB Reconstruction Network (G_A) across multiple scales. For the preservation and enhancement of the rich textural information intrinsic to NIR inputs, our approach utilizes the Laplacian operator within the Fusion Block. This step is crucial for isolating texture features from the NIR images, which are then adeptly combined with the color map produced by sub-network G_B, employing the SPADE Resnet Block (Sun and Jung, 2020)—a process that notably augments the accuracy of color information in distinct regions. Lastly, to achieve a harmonious integration of the NIR feature maps and texture-enriched HSV color maps, we incorporate a plug-and-play cross-attention block (Huang et al., 2019), a strategic move that promotes the seamless convergence of the color predictions generated by generator G_A with the intricate texture-enriched HSV color maps. This methodical fusion paves the way to the final output, with subsequent updates to the discriminator and generator to refine the translation process.

We will introduce details of the RGB Reconstruction Network, HSV Color Prediction Sub-network, Visual State Space Block, and Objectives in the following sections.

3.3. RGB Reconstruction Network

We utilize a U-net (Ronneberger et al., 2015) structure for this dense prediction task. The encoder phase of the network, as depicted in Figure 3 **(a)**, integrates a sequence of downsampling layers, each incorporating our specially devised Visual State Space Block (VSSB).

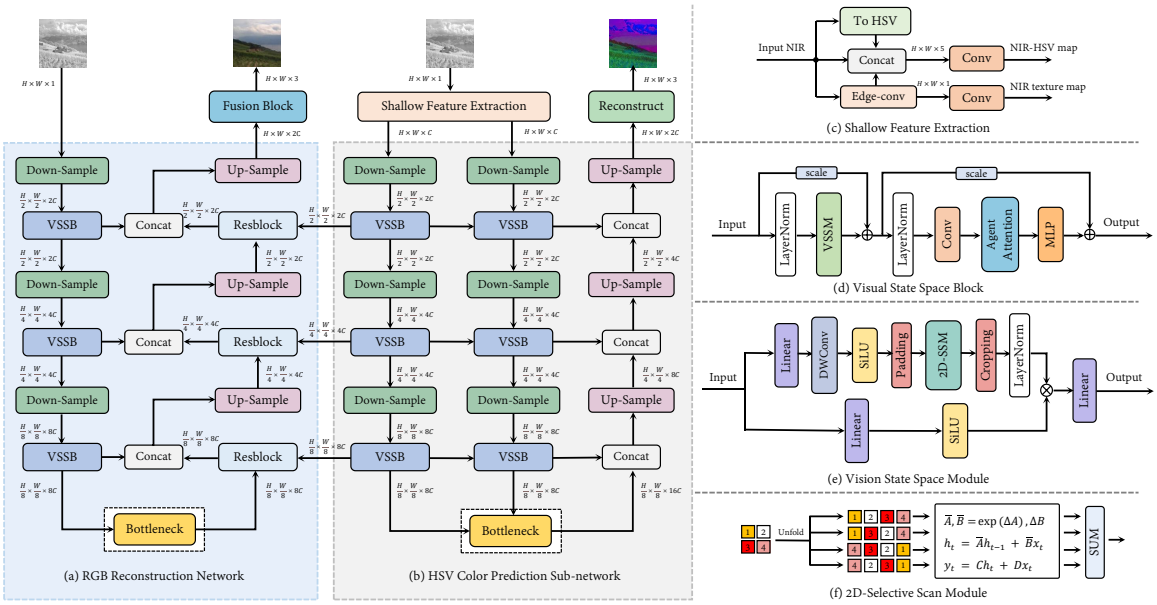

Figure 3: **The pipeline of ColorMamba.** The model consists of two generative networks: **(a)** the RGB Reconstruction Network (G_A) and **(b)** the HSV Color Prediction Sub-network (G_B). **(c)**, **(d)**, **(e)**, **(f)** illustrate details of the Shallow Feature Extraction layer, Visual State Space block (VSSB), Vision State Space Module (VSSM) and 2D-Selective Scan Module (2D-SSM), respectively.

The VSSB is capable to capture global long-range dependencies and manage local contexts simultaneously, thereby enhancing the efficacy of feature extraction significantly. During the decoding phase, we leverage the SPADE Resnet Block (SRB) to adeptly amalgamate multi-scale features derived from the HSV Color Prediction Sub-network. The SRB is adept at aligning the intermediate features with their respective scales, substantially improving the color authenticity of the resulting colorized NIR images and mitigating color distortions. After the decoding stage, the synthesized output is fed into the fusion module for further refinement.

3.4. HSV Color Prediction Sub-network

Recognizing the critical role of accurate color translation from RGB references to NIR images, we have devised an innovative HSV Color Prediction Sub-network, as illustrated in Figure 3 **(b)**. This sub-network processes a monochromatic NIR image ($x_{nir} \in \mathbb{R}^{H\times W\times 1}$) and is trained to mimic the coloration patterns found in ground truth RGB images. Our research indicates that the HSV color space effectively represents these patterns. The objective is to produce an HSV image ($y_{hsv} \in \mathbb{R}^{H\times W\times 3}$) that accurately mirrors the true colors of the RGB image, thereby providing a robust color prior for the main RGB Reconstruction Network.

The process begins with the NIR inputs passing through the Shallow Feature Extraction (SFE) module, outlined in Figure 3 **(c)**. Within the SFE, the NIR image is subject to three distinct operations:

(i) Expansion and Transformation, which augments the image to three channels and transitions it into the HSV space ($x_{hsv} \in \mathbb{R}^{H\times W\times 3}$);

(ii) Texture Feature Extraction, wherein texture patterns are identified and processed to create a texture map;

(iii) Feature Concatenation, involving the concatenation of the original NIR image, the transformed X_{hsv}, and the extracted texture features, resulting in a composite NIR-HSV texture map. This additional texture detail ensures that color edges remain distinct throughout the HSV learning phase, thus reducing the chances of color bleeding.

The whole process of SFE is expressed as follows:

$$\begin{aligned} X_{hsv} &= \text{To-HSV}(X_{nir}), \\ X_{edge} &= \text{Edge-conv}(X_{nir}), \\ X_{tex} &= \text{Conv}(X_{edge}), \\ X_{nir\text{-}hsv} &= \text{Conv}(\text{Concat}(X_{nir} + X_{edge} + X_{hsv})) \end{aligned} \tag{4}$$

The main structure of our HSV Color Prediction Sub-network is similar to the RGB Reconstruction Network. Finally, the generated HSV color map y_{hsv}, calibrated against the transformed ground truth HSV values derived from RGB images. is imbued with rich color data and serves as the foundation for the subsequent fusion processes.

3.5. Visual State Space Block

At the heart of ColorMamba lies our innovative Visual State Space Block (VSSB), designed specifically for capturing extensive spatial relationships while simultaneously amplifying the local context within images. As depicted in Figure 3 **(d)**, the VSSB integrates several critical elements, including the Vision State Space Module (VSSM) and agent-based attention mechanisms.

The process begins by normalizing input deep features ($X \in \mathbb{R}^{H\times W\times C}$) with LayerNorm. Subsequently, we apply the VSSM to extract long-range spatial dependency and local contextual features. The VSSM utilizes a combination of linear operations, convolutional layers, activation functions, and a 2D-Selective Scan Module (2D-SSM) to adeptly gather spatial details from the input features. Moreover, we also use the learnable scale factor ($s \in \mathbb{R}^{C}$) to control the information from skip connection.

A conventional 2D-SMM, as used in the standard Mamba (Gu and Dao, 2023), typically processes images by converting them into one-dimensional sequences for recursive computations. This method risks disrupting the spatial continuity of neighboring pixels, leading them to occupy remote positions in the sequence and resulting in what is known as ***local context neglect***—a loss of the natural spatial correlation between adjacent pixels, especially in the distinction of image boundaries.

To rectify this tendency towards contextual oversight, we have designed a novel scanning methodology. It involves injecting learnable padding tokens between non-adjacent tokens in the state space sequence where there lack of direct spatial correlation. This insertion of padding effectively assists Mamba blocks in a more precise delineation of image borders,

Algorithm 1 Training Strategy of ColorMamba

Require: NIR input image set A, RGB ground-truth image set B, the number of Generator iterations per generator iteration n_{gen}, batch size m, and the number of epoch n_e
Require: Randomly initialize generator parameters θ_g, and discriminator parameters θ_d
for $k = 1, 2, ..., n_e$ **do**
  Sample m NIR images $\{a^{(i)}\}_{i=1}^{m}$ from A
  Sample m RGB images $\{b^{(i)}\}_{i=1}^{m}$ from B
  Obtain colorized NIR data: $\{G_A(a^{(i)})\}_{i=1}^{m}$
  Obtain paired HSV data: $\{G_B(a^{(i)})\}_{i=1}^{m}$
  Obtain NIR texture data: $\{Lap(a^{(i)})\}_{i=1}^{m}$
  Fusion of $\{G_A(a^{(i)})\}_{i=1}^{m}$, $\{G_B(a^{(i)})\}_{i=1}^{m}$ and $\{Lap(a^{(i)})\}_{i=1}^{m}$ to obtain enhanced data: $\{F(G(a^{(i)}))\}_{i=1}^{m}$
  Update the discriminator: $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\lambda_{adv} \log D(b^{(i)}) + \lambda_{adv} \log(1 - D(F(G(a^{(i)}))))\right]$
  for $t = 1, 2, ..., n_{gen}$ **do**
    Sample m NIR images $\{\boldsymbol{a}^{(i)}\}_{i=1}^{m}$ from A
    Sample m RGB images $\{\boldsymbol{b}^{(i)}\}_{i=1}^{m}$ from B
    Obtain colorized NIR data: $\{G_A(\boldsymbol{a}^{(i)})\}_{i=1}^{m}$
    Obtain paired HSV data: $\{G_B(\boldsymbol{a}^{(i)})\}_{i=1}^{m}$
    Obtain NIR texture data: $\{Lap(\boldsymbol{a}^{(i)})\}_{i=1}^{m}$
    Fusion of $\{G_A(\boldsymbol{a}^{(i)})\}_{i=1}^{m}$, $\{G_B(\boldsymbol{a}^{(i)})\}_{i=1}^{m}$ and $\{Lap(\boldsymbol{a}^{(i)})\}_{i=1}^{m}$ to obtain enhanced data: $\{F(G(\boldsymbol{a}^{(i)}))\}_{i=1}^{m}$
    Update the generator: $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \left[\lambda_{adv} \log(1 - D(F(G(\boldsymbol{a}^{(i)})))) + \lambda_{mse}\mathcal{L}_{\text{mse}} + \lambda_{fea}\mathcal{L}_{\text{fea}}\right]$
  end for
end for

enhancing the model's spatial discernment and maintaining the integrity of the sequential representation. Details are provided in the **Preliminaries** 3.1 and illustrated in Figure 2. The above process is expressed as follows:

$$\begin{aligned} X_1 &= \text{LN}(\text{Cropping}(\text{2D-SSM}(\text{Padding}(\text{SiLU}(\text{DWConv}(\text{Linear}(\text{LN}(X_{in}))))))))), \\ X_2 &= \text{SiLU}(\text{Linear}(\text{LN}(X_{in}))), \\ X_3 &= \text{Lincar}(X_1 \odot X_2) + s \cdot X_{in} \end{aligned} \tag{5}$$

where DWConv represents depth-wise convolution, and $\odot$ denotes the Hadamard product. Moreover, by incorporating localized convolutional enhancements and a plug-and-play agent attention (Han et al., 2023), we enhance the model's capacity for local context recognition. These improvements transform the conventional Mamba block into the more advanced VSSB, enabling our model to proficiently map both the overarching spatial interdependencies and the finer, localized contextual nuances within the image data. Another adjustable scaling factor ($s' \in \mathbb{R}^C$) is used in residual connection, and the process can be expressed as:

$$\begin{aligned} X_4 &= \text{MLP}(\text{Agent}(\text{Conv}(\text{LN}(X_3)))), \\ X_{out} &= X_4 + s' \cdot X_3 \end{aligned} \tag{6}$$

3.6. Objectives

We provide a pseudo-code to depict the whole training process of our model, as shown in Algorithm 1. Specifically, we use three loss functions to formulate our final objective:

MSE Loss: We use MSE loss as pixel-wise supervision between each predicted value x_i and its corresponding ground truth y_i:

$$\mathcal{L}_{\text{mse}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \tag{7}$$

Table 1: **Quantitative comparison**. The best results are highlighted in bold.

Methods	PSNR(↑)	SSIM(↑)	AE (↓)	LPIPS(↓)	SAM (↓)	ERGAS(↓)
SST (Yan et al., 2020)	14.26	0.57	5.61	0.361	0.147	15.32
NIR-GNN (Valsesia et al., 2020)	17.50	0.60	5.22	0.384	0.113	13.27
MFF (Yan et al., 2020)	17.39	0.61	4.69	0.318	0.106	12.85
ATCGAN (Yang and Chen, 2020)	19.59	0.59	4.33	0.295	0.085	9.42
Restormer (Zamir et al., 2022)	19.43	0.54	4.41	0.267	0.077	8.07
DRSformer (Chen et al., 2023a)	20.18	0.56	4.22	0.254	0.074	8.04
MPFNet (Yang et al., 2023b)	22.14	0.63	3.68	0.253	0.067	6.54
CoColor (Yang et al., 2023a)	<u>23.54</u>	<u>0.69</u>	**2.68**	0.233	<u>0.059</u>	<u>5.41</u>
MCFNet (Zhai et al., 2024)	20.34	0.61	3.79	**0.208**	0.083	7.94
ColorMamba(ours)	**24.56**	**0.71**	<u>2.81</u>	<u>0.212</u>	**0.049**	**3.27**

Feature Consistency Loss: We further introduce a perceptual loss based on a pretrained autoencoder (Ng et al., 2011), which combines MSE, Cosine Similarity, and Multi-Scale Structural Similarity (MS-SSIM) index:

$$\mathcal{L}_{\text{fea}} = \alpha\mathcal{L}_{\text{mse}}(X,Y) + \gamma\mathcal{L}_{\text{cosine}}(X,Y) + \beta\mathcal{L}_{\text{ms-ssim}}(X,Y) \tag{8}$$

Adversarial Loss: The adversarial loss is defined as follows:

$$\mathcal{L}_{\text{adv}}(G,D,X,Y) = \mathbb{E}_{Y\sim p_{data}(Y)}[\log D(Y)] + \mathbb{E}_{X\sim p_{data}(X)}[\log(1-D(G(X)))] \tag{9}$$

Full Objective Function: The total loss can be expressed as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{mse}\mathcal{L}_{\text{mse}} + \lambda_{fea}\mathcal{L}_{\text{fea}} + \lambda_{adv}\mathcal{L}_{\text{adv}} \tag{10}$$

where λ_{mse}, λ_{fea}, and λ_{adv} are hyperparameters to balance weights of different terms.

4. Evaluation

In this section, we will first introduce the implementation details of our model. Next, we will evaluate our framework by presenting quantitative and qualitative results on the spectral translation task and comparing it with other state-of-the-art methods. Finally, we will validate the effectiveness of the proposed framework through ablation studies.

4.1. Implementation Details

We use the VCIP2020 Grand Challenge on the NIR dataset (Yang et al., 2023a) to train and test our network. Data augmentation techniques (Yang and Chen, 2020) are employed, including random resizing, cropping, contrast adjustment, and image mirroring. All images are within a resolution of 256×256, and are normalized to the range $(0, 1)$ during the training process. We use ResNet as the backbone network with an initial learning rate of 1e-4. The network is trained using the AdamW optimizer (Loshchilov and Hutter, 2017), with parameters set to $\beta_1 = 0.5$, $\beta_2 = 0.999$, and weight decay $= 0.5$. For the parameters in the loss function, we set $\lambda_{mse} = 100$, $\lambda_{fea} = 100$ and $\lambda_{adv} = 1$. The entire network is trained end-to-end in a self-supervised manner for 300 epochs, with a batch size of 8.

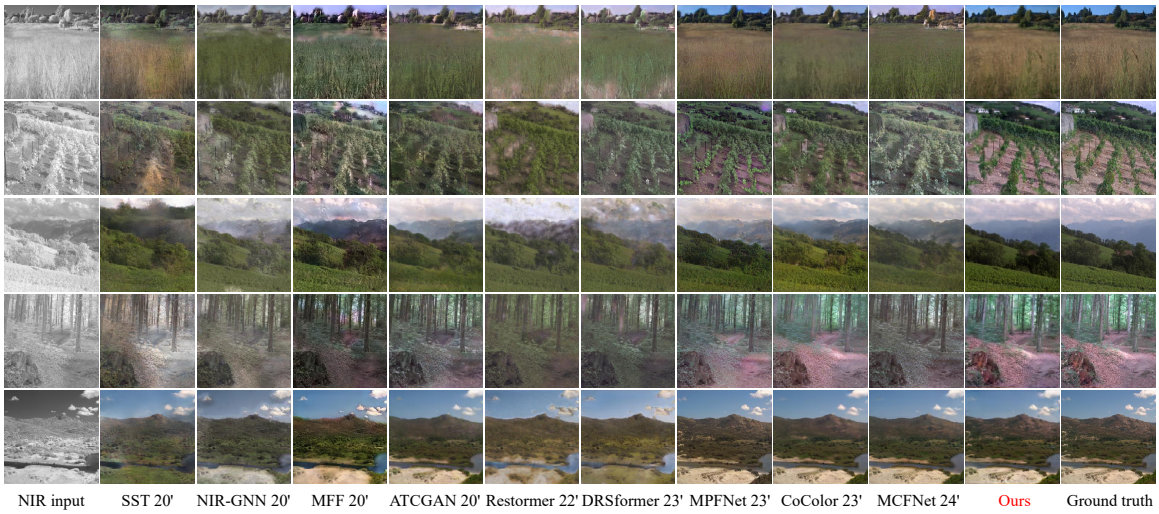

Figure 4: **Visual comparison** of different methods on testing datasets. From left to right are SST (Yan et al., 2020), NIR-GNN (Valsesia et al., 2020), MFF (Yan et al., 2020), ATCGAN (Yang and Chen, 2020), Restormer (Zamir et al., 2022), DRSformer (Chen et al., 2023a), MPFNet (Yang et al., 2023b), CoColor (Yang et al., 2023a), and MCFNet (Zhai et al., 2024).

4.2. Comparison Experiments

We quantitatively and qualitatively compared our MCFNet method with 7 spectral translation methods, including SST (Yan et al., 2020), NIR-GNN (Valsesia et al., 2020), MFF (Yan et al., 2020), ATCGAN (Yang and Chen, 2020), MPFNet (Yang et al., 2023b), CoColor (Yang et al., 2023a), and MCFNet (Zhai et al., 2024), and 2 image restoration methods, including Restormer (Zamir et al., 2022) and DRSformer (Chen et al., 2023a), to demonstrate the performance of our method.

Quantitative Evaluation: For the performance assessment of our model, we utilized six metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Absolute Error (AE), Learned Perceptual Image Patch Similarity (LPIPS), Spectral Angle Mapper (SAM), and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS).

As shown in Table 1, our model outperforms other methods in terms of PSNR, SSIM, SAM, and ERGAS measures. Particularly noteworthy are the PSNR and ERGAS outcomes, where ColorMamba exhibits substantial leverage, with a 1.02 improvement in PSNR and nearly a 40% improvement in ERGAS, corroborating its capacity to produce colorization results that approach naturalistic and credible visual qualities.

Qualitative Evaluation: We visualize the spectral translation results in Figure 4. As can be seen, SST (Yan et al., 2020), NIR-GNN (Valsesia et al., 2020), MFF (Yan et al., 2020) and ATCGAN (Yang and Chen, 2020) all failed to recover the vivid color distribution of RGB ground truths and retain the texture details of NIR inputs. Restormer (Zamir

Table 2: **Ablation studies on ColorMamba.** The best results are highlighted in bold.

Variants	PSNR(↑)	SSIM(↑)	AE (↓)	LPIPS(↓)	SAM (↓)	ERGAS(↓)
w/o Mamba	24.25	0.70	2.88	0.242	0.050	**3.20**
w/ Mamba	23.97	0.68	2.88	0.250	0.052	3.50
w/ Mamba+Att	24.36	0.70	2.82	0.220	**0.049**	3.35
w/ Mamba+Att+padding	**24.56**	**0.71**	**2.81**	**0.212**	**0.049**	3.27

et al., 2022) and DRSformer (Chen et al., 2023a) exhibit limited capacity in the spectral translation task, often resulting in images that are dull, lackluster, and rife with color inaccuracies when compared to authentic imagery. CoColor (Yang et al., 2023a) handles complex scenarios with commendable color consistency, particularly evident in the fourth row. However, its output is deficient in terms of sharpness and detail, with shortcomings noticeable in the second and fifth rows, along with occasional color variance in the first and third rows. In contrast, our approach stands out among these methods, delivering accurate color restoration that convincingly approximates real-life imagery. This success is largely attributed to the Visual State Space Block (VSSB), which captures long-range dependencies and local contextual features efficiently.

4.3. Ablation Experiments

Our ColorMamba architecture is distinguished by the integration of standard Mamba (Gu and Dao, 2023), local context enhancement (*e.g.*, agent-based attention (Han et al., 2023) mechanisms), and learnable padding token injection. Central to its functionality is the Visual State Space Block (VSSB), which is pivotal for spectral translation — crucial for preserving spatial coherence and ensuring precise spectral data representation. To evaluate the contribution of individual components within the ColorMamba framework, we engaged in a series of ablation experiments. The objective of these experiments was to methodically deconstruct the model by selectively deactivating or modifying specific elements and examining the subsequent effects on performance. The findings in Table 2 reveal that:

(i) Omitting the use of the Mamba block impairs the model's performance, as it results in a deficiency in positional context during the translation between different domains.

(ii) The deployment of the standard Mamba block without adaptations yields inferior outcomes since its direct application to two-dimensional vision tasks overlooks the nuances of local contexts.

(iii) The learnable padding tokens enable the Mamba block to interpret edge regions more precisely. This enhances both spatial awareness within the model and the cohesiveness of sequential data handling. Neglecting to use these padding tokens leads to the generation of merely subpar results.

5. Limitation

Spectral translation is an ill-posed problem due to its cross-domain nature. The challenge arises due to both intensity and chrominance need to be estimated, which is required by the disparate spectral bands characteristic of the NIR and visible spectra. Further complexity is introduced by environmental variations, such as thermal changes and alterations in the

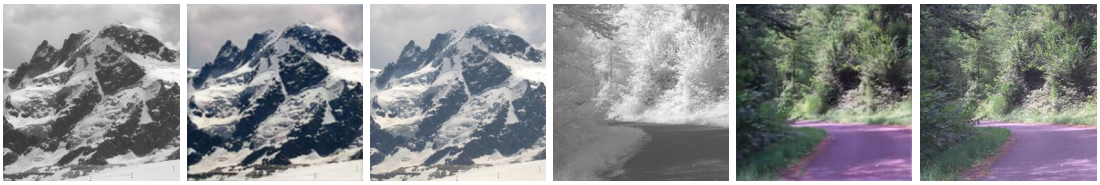

Figure 5: **Visual examples of deficiencies.** Our ColorMamba generates some oversaturated images compared to ground truths.

light source, which can lead to substantial variances in the intensity of NIR images captured even within identical settings of the training dataset, thereby exacerbating the ambiguity in mapping (Yang et al., 2023b). This phenomenon aggravates the extraction and integration of color information from the HSV model, and thus leads to some over-saturated results, as shown in Figure 5.

6. Conclusion

In the present study, we have introduced a novel spectral translation framework termed ColorMamba. This innovative model integrates enhancements in local convolution, attention mechanisms, and a pioneering scanning technique within its foundational Visual State Space Blocks (VSSBs). These enhancements collectively facilitate a more nuanced exploration of both extensive long-range relationships and detailed local context within the spectral translation domain. Moreover, the application of HSV color priors furnishes multi-scale guidance through the reconstruction phase, culminating in a more precise spectral translation. A comprehensive suite of experimental evaluations confirms that our proposed method surpasses existing baseline methodologies in performance, suggesting that our ColorMamba offers a potent and auspicious foundational architecture for endeavors in spectral translation.

References

Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023a.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023b.

F Christnacher, E Bacher, N Metzger, S Schertzer, Y Lutz, J-M Poyet, and M Laurenzis. Portable bi-λ SWIR/NIR GV gated viewing system for surveillance and security applications. In *Electro-Optical Remote Sensing XII*, pages 54–64, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain

Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Scalable diffusion models with state space backbone. *arXiv preprint arXiv:2402.05608*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021a.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021b.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024.

Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. *arXiv preprint arXiv:2312.08874*, 2023.

Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer. Zigma: Zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802*, 2024.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Armin Mehri and Angel D Sappa. Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241, 2015.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Shubhabrata Sengupta, Mark J Harris, Michael Garland, and John D Owens. *Efficient parallel scan algorithms for many-core gpus*. eScholarship, University of California, 2011.

Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020.

Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–23, 2017.

Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Learning to colorize infrared images. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15*, pages 164–172, 2018.

Tian Sun and Cheolkon Jung. Nir image colorization using spade generator and grayscale approximated self-reconstruction. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 463–466, 2020.

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43, 2017.

Prasad S Thenkabail, John G Lyon, and Alfredo Huete. Advances in hyperspectral remote sensing of vegetation and agricultural crops. In *Fundamentals, Sensor Systems, Spectral Libraries, and Data Mining for Vegetation*, pages 3–37. 2018.

Diego Valsesia, Giulia Fracastoro, and Enrico Magli. NIR image colorization with graph-convolutional neural networks. In *IEEE VCIP*, pages 451–454, 2020.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023.

Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8239–8249, 2024.

Longbin Yan, Xiuheng Wang, Min Zhao, Shumin Liu, and Jie Chen. A multi-model fusion framework for NIR-to-RGB translation. In *IEEE VCIP*, pages 459–462, 2020.

Xingxing Yang, Jie Chen, and Zaifeng Yang. Cooperative colorization: Exploring latent cross-domain priors for nir image spectrum translation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2409–2417, 2023a.

Xingxing Yang, Jie Chen, and Zaifeng Yang. Multi-scale progressive feature embedding for accurate nir-to-rgb spectral domain translation. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2023b.

Xingxing Yang, Jie Chen, and Zaifeng Yang. Hyperspectral image reconstruction via combinatorial embedding of cross-channel spatio-spectral clues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6567–6575, 2024.

Zaifeng Yang and Zhenghua Chen. Learning from paired and unpaired data: Alternately trained cyclegan for near infrared image colorization. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 467–470, 2020.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

Huiyu Zhai, Mo Chen, Xingxing Yang, and Gusheng Kang. Multi-scale hsv color feature embedding for high-fidelity nir-to-rgb spectrum translation. *arXiv preprint arXiv:2404.16685*, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.