CO^{2}IF: Language-Bridging Hyperspectral-Multispectral Image Fusion with Coordinated and Cross-modal Optimal Transport

Mingjin Zhang, Zhongkai Yang*, Fei Gao*

Xidian University, China

mjinzhang@xidian.edu.cn, 23011210500@stu.xidian.edu.cn, fgao@xidian.edu.cn

Abstract

Due to the difficulties of directly obtaining *high-resolution hyperspectral images* (HR-HSI), the fusion of *low-resolution hyperspectral images* (LR-HSI) and *high-resolution multispectral images* (HR-MSI) has emerged as an effective approach. While existing methods leverage image-level priors from HR-MSI, they often lack explicit semantic guidance for precise detail reconstruction. Recognizing that textual scene descriptions encapsulate valuable object attributes and contextual information, we introduce the first *Language-Bridging framework for Hyperspectral and Multispectral image fusion* (CO^{2}IF). CO^{2}IF leverages language semantics as prior knowledge to explicitly guide the reconstruction process. To bridge the modality gap between textual descriptions and high-dimensional hyperspectral data, we design a *Cross-modal Optimal Transport* (COT) module. COT establishes precise semantic correspondences between language features and the visual cues of individual spectral bands. Building upon this semantic alignment, we develop a *Multimodal Coordinated State Space Model* (CoMamba). CoMamba effectively integrates the language-derived priors with spatial information from HR-MSI and spectral information from LR-HSI. This language-guided reconstruction significantly enhances the extraction of crucial spatial-spectral details, leading to superior fidelity in the generated HR-HSI. In addition, this paper adds text descriptions for three widely used datasets. Both qualitative and quantitative experimental results on the public datasets confirm the superiority of the proposed method compared to the SOTA methods.

Code — https://github.com/yzkLearning/CO2IF

Introduction

Hyperspectral imaging sensors can simultaneously capture dozens or even hundreds of spectral bands, covering wavelengths from the visible to the shortwave infrared range, resulting in *hyperspectral images* (HSI) with rich spectral information (Xiao and Wei 2023). Due to the varying reflectance properties of different materials, HSIs can discern subtle spectral differences between them. This makes them valuable in diverse applications, including image classification (Tang et al. 2024) and face recognition (Zhang et al. 2019a,b). With the rise of the medical metaverse, HSIs provide the unique spectral data essential for constructing ultra-realistic virtual tissues, capturing subtle variations that enhance diagnostic precision (Wang et al. 2022; Bashir et al. 2023). However, limited by various hardware in practice, hyperspectral imaging still faces a significant challenge: the trade-off between spatial resolution and spectral resolution (Loncan et al. 2015; Yan et al. 2025). Hyperspectral and Multispectral Image Fusion (HS/MS fusion) algorithms have gained significant attention. They generate *high-resolution hyperspectral images* (HR-HSI) by combining the spatial information of *high-resolution multispectral images* (HR-MSI) with the spectral information of *low-resolution hyperspectral images* (LR-HSI).

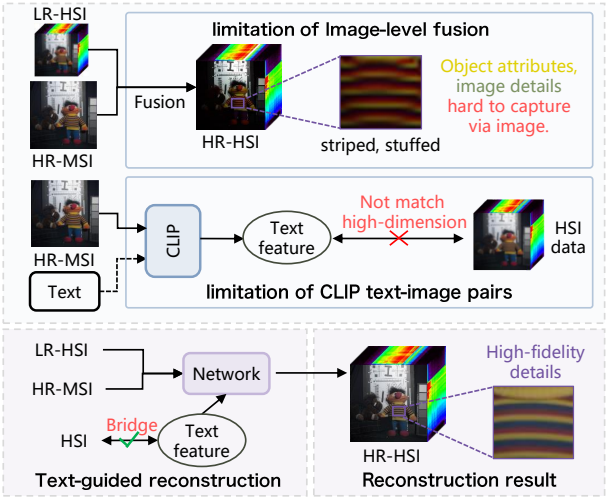

Figure 1: The limitations of previous methods and the advantages of the proposed text-guided reconstruction method are illustrated. Text description provides valuable semantics, which helps to capture high-fidelity detail information during the reconstruction process.

HS/MS fusion methods can be divided into traditional algorithms and deep learning-based ones. Early methods (Aiazzi et al. 2006; Chen et al. 2014; Dong et al. 2016) utilize handcrafted models and prior knowledge to analyze

*Corresponding author

intrinsic relationships within hyperspectral data. However, they often failed to effectively capture spatial-spectral correlations across diverse scenarios. The current mainstream deep learning methods are primarily based on Convolutional Neural Networks (CNNs) and Transformers have emerged (Han et al. 2020; Qu, Qi, and Kwan 2018; Peng et al. 2023). Most existing methods extract features from LR-HSI and HR-MSI separately and then fuse them to generate HR-HSI. Leveraging their powerful end-to-end learning capabilities, these methods outperform traditional approaches.

Despite certain progress, methods relying solely on image-modal inputs often lack explicit semantic guidance to enhance object critical features, as shown in Figure 1. Therefore, we propose leveraging text descriptions that present scene styles and attributes of main objects to provide explicit semantic guidance for feature recovery. However, integrating text into reconstruction networks poses two challenges: (1) **Feature alignment.** Vision-language models are trained on RGB images, while hyperspectral data differs notably from such image data. Thus, how can we improve the alignment between hyperspectral features and text information? (2) **Guidance of text descriptions.** Due to feature differences across modalities, how can we utilize the obtained image-level priors and text semantic information to accurately guide HSI reconstruction?

To this end, we propose the first *Language-Bridging framework for Hyperspectral and Multispectral image fusion* (CO^2IF) in this paper. First, we introduce a *Cross-modal Optimal Transport* (COT) module, which models the alignment process as a transportation problem, thereby achieving accurate cross-modal mapping between language and hyperspectral data. By minimizing the cost matrix, it assigns local semantics to the visual cues of each spectral band within a global context, thus realizing the dual constraints of global scene consistency and band-specific enhancement. On this basis, we propose a *Mutlimodal Coordinated State Space Model* (CoMamba) that leverages Mamba's competitive contextual information modeling capabilities at linear complexity. It capitalizes on the homogeneity of identical objects in the background, effectively enhancing the extraction of key visual cues across spectral bands while minimizing redundancy. Additionally, we extend the public CAVE, Chikusei, and Harvard datasets with corresponding textual data. The text descriptions are generated by GPT-4V and manually reviewed for accuracy. Experimental results validate the advantages of CO^2IF compared to SOTA methods.

In summary, our main contributions are as follows:

- We propose CO^2IF, the first vision-language framework with language-bridging prior knowledge learning for HS/MS fusion task.
- We develop two key components, *i.e.* COT and CoMamba, to link semantics with hyperspectral data and extract deep features across modalities.
- We extend the CAVE, Chikusei, and Harvard datasets with corresponding text data, providing rich resources for multimodal HS/MS fusion.
- Extensive experiments demonstrate that the proposed CO^2IF surpasses SOTA methods, excelling in both objective metrics and visual quality.

Related Work

Hyperspectral and Multispectral Image Fusion. Traditional methods utilize handcrafted priors to integrate MSI spatial information into HSI. Bayesian enable tailored prior distributions for regularization (Akhtar, Shafait, and Mian 2015). Unmixing techniques decompose inputs into bases and combine them with HR-MSI for reconstruction (Dian et al. 2019). With the advancement of neural networks (Zhang et al. 2022a,b,c), deep learning methods leverage CNNs and Transformers to extract spatial-spectral features more effectively. For example, BDT (Deng et al. 2023) designs a bidirectional dilation Transformer that leverages latent multi-scale information to generate high-quality target HSI. HSR-Diff (Wu et al. 2023a) applies the conditional diffusion model, using a conditional denoising Transformer to eliminate noise during iterative refinement. FusionMamba (Peng et al. 2024) extracts features from LR-HSI and HR-MSI through two U-shaped structures composed of Mamba. While these methods show promising performance, their reliance on image modalities limits the capture of contextual information, restricting visual representation across spectral bands. Language mode can provide homogeneous feature representations, aiding in the extraction of detailed features from high-dimensional hyperspectral data.

Vision-Language Models. Recently, the development of foundational models has entered a new era, with a notable example being Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) for vision-language cross-modal tasks. CLIP is pretrained on a large-scale image-text pair dataset, providing rich semantic knowledge for vision, and has demonstrated remarkable zero-shot performance across various computer vision tasks (Zhong et al. 2022; Wu et al. 2023b; Weng et al. 2023). Notably, recent work has successfully applied CLIP to natural image super-resolution (Hu et al. 2024a). However, as CLIP models are tailored for RGB imagery, they fail to accommodate the high-dimensional nature of hyperspectral data and thus cannot be directly applied to hyperspectral/multispectral (HS/MS) fusion. To address this limitation, we integrate CLIP's semantic knowledge with hyperspectral information, leveraging this consistency to guide high-fidelity HSI reconstruction with enhanced fine-grained details. This demonstrates the feasibility and effectiveness of incorporating language priors into HS/MS fusion tasks.

Selective State Space Models (Mamba). Selective State Space Models, which utilize state and observation equations for system modeling, combine linear computational complexity with competitive contextual information capture capabilities. The vision model combined with Mamba (Gu and Dao 2023) has been successful in various computer vision tasks. Vim (Zhu et al. 2024) leverages bidirectional scanning to enhance visual token processing in both forward and backward directions. VMambaSCI (Zhang et al. 2024) introduces a dual-domain scanning Mamba, successfully applied to compressed spectral imaging. MambaIRv2 (Guo et al. 2024) explores the connection between Mamba and linear

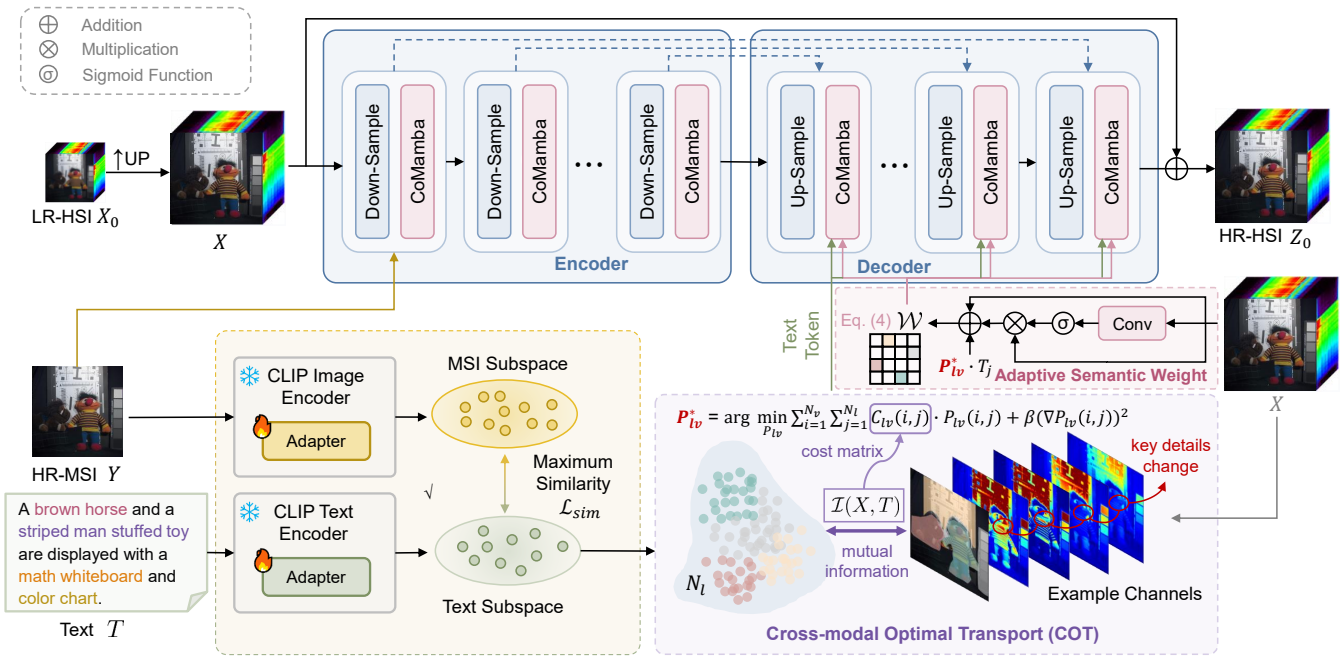

Figure 2: Overview of the CO^{2}IF for hyperspectral and multispectral image fusion. The framework reconstructs high-quality HR-HSI through a sequential, cross-modal aligned process: (1) Upsampled LR-HSI and HR-MSI are fed into the encoder to extract initial spectral and spatial features. (2) Align text and HR-MSI via contrastive loss in CLIP encoder to bridge semantic-visual gaps. (3) Text tokens match upsampled LR - HSI through COT for optimal correspondence. (4) Feed text tokens and adaptive weight $\mathcal{W}$ into CoMamba in the decoder to generate high-quality HR-HSI.

attention, enabling Mamba to exhibit non-causal modeling capabilities similar to ViTs. Multimodal mamba, such as Cobra (Zhao et al. 2024) and VL-Mamba (Qiao et al. 2024), are applied to vision-language tasks. They primarily concatenate visual and textual features, which are subsequently fed into the Mamba framework for sequential processing. However, due to the high-dimension of HSIs, they struggle to capture the complex relationship between text and HSI.

Method

Overview

The overall architecture is illustrated in Figure 2. It introduces language-derived priors to assist the fusion of spatial and spectral information between LR-HSI and HR-MSI, alleviating potential artifacts and distortions, and generating high-quality HR-HSI. Specifically, text and HR-MSI pairs are first fine-tuned through a CLIP encoder to achieve maximum similarity, where the adapter adopts a linear-GELU-linear structure. Subsequently, the text is processed by the COT, leveraging an optimal transport method with mutual information to achieve the optimal mapping between text features and the key visual cues of each hyperspectral band. The optimal transport matrix generated by COT is further utilized to compute an adaptive semantic weight matrix tailored to spectral characteristics. This matrix, along with the aligned text tokens, is fed into the CoMamba in the decoder, providing rich semantics for the HSI reconstruction process.

Cross-modal Optimal Transport (COT)

The challenge of HSI data fusion is how to effectively preserve the spectral information of LR-HSI and improve the spatial resolution. To address this issue, we further process the text tokens aligned by the CLIP encoder (Hu et al. 2024b) to achieve complex alignment between language features and the features of different spectral bands in HSI. This alignment helps the network effectively preserve detailed visual cues across various bands, enhancing both spectral and spatial representations. As shown in In Figure 2, we propose COT—an advanced optimization-theory-based mathematical strategy. It models the mapping process as a regularized optimal transport problem (Xu et al. 2020).

Cross-modal Optimal Transport Process. Given feature sequences $X \in R^{N_v \times B}$ and $T \in R^{N_l \times B}$ derived from the HSI modality and the text feature sequence obtained from the CLIP encoder, respectively. Here, N_v and N_l represent sequence lengths for different modalities, and B stands for the channel dimension. Our objective is to learn an optimal transport matrix P_{lv} that models fine-grained correspondences between text and HSIs. The optimal transport problem can be formulated:

$$\min_{P_{lv}} \sum_{i=1}^{N_v}\sum_{j=1}^{N_l} C_{lv}(i,j) \cdot P_{lv}(i,j) + \beta(\nabla P_{lv}(i,j))^2, \quad (1)$$

where $C_{lv} \in R^{N_v \times N_l}$ is the cost matrix, with N_v and N_l both derived from the output of CLIP. The regular term

$(\nabla P_{lv}(i,j))^2$ controls the smoothness of the map, helping to speed up convergence while also avoiding numerical instability. β is a hyperparameter obtained through training.

Cross-modal Mutual Information. Considering the possibility of fuzzy rough matching in the optimal transmission process, we apply mutual information (Kraskov, Stögbauer, and Grassberger 2004) to quantify the correlation between feature vectors, thereby reducing erroneous pairwise transport. The mutual information between text and HSI features can be described as:

$$\begin{aligned}\mathcal{I}(X,T) &= \mathcal{H}(X) + \mathcal{H}(T) - \mathcal{H}(X,T) \\ &= -\sum_{X_i \in N_v} p(X_i)\log p(X_i) - \sum_{T_j \in N_l} p(T_j)\log p(T_j) \\ &\quad + \sum_{X_i \in N_v}\sum_{T_j \in N_l} p(X_i,T_j)\log p(X_i,T_j),\end{aligned} \tag{2}$$

where $\mathcal{H}(X)$ and $\mathcal{H}(T)$ are the Shannon entropy of HSI and text, respectively, and $\mathcal{H}(X,T)$ stands for the joint entropy. $p(\cdot)$ represents the marginal probability distribution and $p(X_i,T_j)$ denotes joint probability distribution.

Cost Matrix. We incorporate the mutual information $\mathcal{I}(X,T)$ into the cost matrix C_{lv} via cosine similarity:

$$\begin{aligned}C_{lv}(i,j) &= (1 - \frac{X_i \cdot T_j}{\|X_i\|_2 \|T_j\|_2}) \cdot (1 - \mathcal{I}(X,T)) \\ &= (1 - \frac{X_i \cdot T_j}{\|X_i\|_2 \|T_j\|_2}) \cdot \mathcal{G}(X,T),\end{aligned} \tag{3}$$

where $\mathcal{G}(X,T)$ is the probability discrimination matrix derived from the feature space structure. This approach constrains the pairing during the transport process such that even if two features are similar, no transport is established unless they satisfy the object class criterion determined by $\mathcal{G}(X,T)$. It ensures precise optimal transport between text and HSIs, while minimizing interference from neighboring objects near the target object's boundaries, thereby providing rich semantic interpretations for the visual cues across different spectral bands of the HSI.

Adaptive Semantic Weight Matrix. Additionally, we hope to leverage the obtained optimal transport matrix $P_{lv}^* \in R^{N_v \times N_l}$ (from Eq. 1 with the Sinkhorn divergence) to generate an adaptive semantic weight matrix, which will further assist in the fusion module described in the next section. It can be calculated as:

$$\mathcal{W} = P_{lv}^* \cdot T + \sigma(\phi_{conv}(X)) \cdot X + X, \tag{4}$$

where σ stands for sigmoid activation function and $\phi_{conv}(\cdot)$ denotes a particular operator composed of convolution layers. With the matrix $\mathcal{W}$, the model can dynamically assign weights according to the spectral characteristics of the target object at each band and spatial position. For instance, some bands may be more important than others in expressing a particular spectral feature of an object, and the calculated weights can help highlight that key information.

Coordinated State Space Model (CoMamba)

After obtaining the optimal mapping of text-hyperspectral image pairs through COT, the text tokens now exhibit maximum similarity with both HR-MSI and LR-HSI. Next, we need to perform cross-modal fusion of the features to preserve key spatial-spectral information and reduce distortion in the reconstructed HSI. However, the general vision Mamba (Liu et al. 2024; Zhang et al. 2024; Zhu et al. 2024) only performs a single image operation on the image modality, cannot leverage semantic clustering information from objects in the hyperspectral scene to assist in LR-HSI reconstruction. Accordling, as illustrated in Figure 3, we design a CoMamba for the decoder.

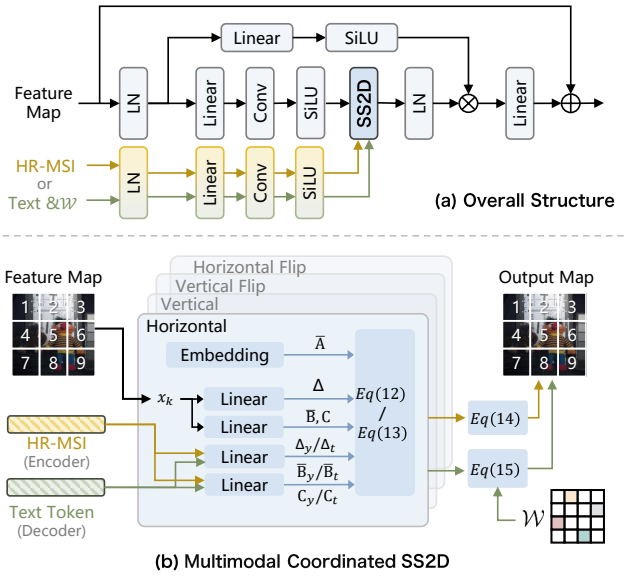

Figure 3: Structure of CoMamba. "LN" denotes layer normalization. The yellow and green arrows represent the encoder and decoder process, respectively.

Vanilla Mamba. The Mamba model (Liu et al. 2024) serves as the foundation of our CoMamba. It is a selective state-space model that relies on a selection mechanism. It transforms an input sequence $x(t) \in R$ into hidden representations $h(t) \in R^H$ and ultimately predicts an output sequence $y(t) \in R$. Mathematically, it can be expressed as:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{5}$$

$$y(t) = \mathbf{C}h(t) + Dx(t), \tag{6}$$

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{7}$$

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \approx \mathbf{B}, \tag{8}$$

where H denotes the size of the state, $\mathbf{A} \in R^{H\times H}$, $\mathbf{B} \in R^H$ and $\mathbf{C} \in R^H$ are learnable parameters, and $D \in R^1$ represents the skip connection. The parameters $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are derived using a timescale parameter Δ discretization.

Multimodal Coordinated SS2D. In CoMamba, we improve its 2D-Selective-Scan (SS2D). We first utilize the {LR-HSI X, HR-MSI Y, text t} triplets to compute the learnable parameters in the Mamba state-space equation. We introduce an additional set of parameters derived from the HR-MSI image (in the encoder) or text (in the decoder), to

augment the original parameter matrices in Mamba. The coordinated parameters can be expressed as:

$$\{\bar{\mathbf{B}}, \Delta, \mathbf{C}\} = \Theta_{proj}(X), \tag{9}$$

$$\text{Encoder:} \quad \{\bar{\mathbf{B}}_y, \Delta_y, \mathbf{C}_y\} = \Theta_{proj}(Y), \tag{10}$$

$$\text{Decoder:} \quad \{\bar{\mathbf{B}}_t, \Delta_t, \mathbf{C}_t\} = \Theta_{proj}(T), \tag{11}$$

where $\Theta_{proj}(\cdot)$ is a projection operator composed of linear norm layers. Δ_t denotes timescale parameter of the text sequence. Then, we update the *hidden representations* by:

$$\text{Encoder:} \quad h_k = \bar{\mathbf{A}} h_{k-1} + (\bar{\mathbf{B}} + \lambda \bar{\mathbf{B}}_y) x_k, \tag{12}$$

$$\text{Decoder:} \quad h_k = \bar{\mathbf{A}} h_{k-1} + (\bar{\mathbf{B}} + \lambda \bar{\mathbf{B}}_t) x_k, \tag{13}$$

and predict the *output sequence* in the encoder by:

$$\text{Encoder:} \quad y_k = (\mathbf{C} + \gamma \mathbf{C}_y) h_k + x_k, \tag{14}$$

where λ and γ are weight influence factors. This allows the model to leverage the high-resolution spatial information of HR-MSI to guide the reconstruction of spatial features.

Semantic-adaptively Weighted Output. In the decoder, we also consider the spectral domain and use the adaptive semantic weight matrix $\mathcal{W}$ obtained in Section to monitor the results of the state-space equation. Specifically, since $\mathcal{W}$ encodes the adaptively adjusted spectral information, it can filter out semantically redundant or unnecessary spectral features, thereby reducing reliance on redundant bands and minimizing artifacts and distortions in the output features. We integrate $\mathcal{W}$ into the state-space equation, updating the output equation as follows:

$$\text{Decoder:} \quad y_k = (\mathbf{C} + \gamma \mathbf{C}_t) h_k \odot \mathcal{W} + x_k, \tag{15}$$

where $\odot$ denotes the pixel-wise multiplication.

Training Objectives

Text-MSI Contrastive Loss. During the training process, we use the contrastive loss (Shi et al. 2019) to maximize the similarity between the text prompt tokens obtained from the CLIP text encoder and the HR-MSI image tokens obtained from the CLIP vision encoder. This alignment helps to integrate the language prior knowledge with the high-resolution spatial information. The contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{sim} = \frac{1}{N} \sum_{i=1}^{N} [& - \log \frac{\exp(y_i^\top t_i)/\tau)}{\sum_{j=1}^{N} \exp(y_i^\top t_j)/\tau} \\ & - \log \frac{\exp(t_i^\top y_i)/\tau}{\sum_{i=1}^{N} \exp(t_i^\top y_j)/\tau}], \end{aligned} \tag{16}$$

where y_i and t_i represent the normalized embeddings of the HR-MSI and text in the i-th pair, respectively. τ is a learned temperature parameter to scale the logits.

Pixel-wise Reconstruction Loss. In addition, we calculate the distance between the network reconstruction HSI Z_0 and the ground truth value Z in terms of L1 losses:

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^{N} \| Z_0 - Z \|_1. \tag{17}$$

Total Loss. The overall loss function is defined as the sum of two loss components:

$$\mathcal{L}_{total} = \mathcal{L}_1 + \mathcal{L}_{sim}. \tag{18}$$

Experiments

Datasets

Cave Dataset: the Cave dataset (Yasuma et al. 2010) is a multispectral dataset that contains full spectral resolution reflectance data from 400 nm to 700 nm at a resolution of 10 nm (31 bands in total), covering 32 images of everyday objects. The dataset was obtained from a Cooled CCD camera and images within have 512×512 pixels and are stored as a 16-bit grayscale image per band. We use 22 randomly selected images for training and the remainder for testing.

Chikusei Dataset: the Chikusei dataset is accessed by the Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan (Yokoya and Iwasaki 2016). The image has 128 bands in the spectral range from 363 nm to 1018 nm. Image of 19 classes is collected using high-resolution images together with the hyperspectral data via a field survey and visual inspection. We crop it to 16 non-overlapping 512×512 size patches, 12 images for training and the rest for testing.

Harvard Dataset: the Harvard dataset (Chakrabarti and Zickler 2011) consists of a total of 50 indoor and outdoor images captured using a commercial hyperspectral camera (Nuance FX, CRI Inc.). Each HSI has a size of $1392 \times 1040 \times 31$ and contains 31 spectral bands ranging from 420 nm to 720 nm, with a wavelength interval of 10 nm. We randomly select 30 images with dimensions of $960 \times 960 \times 31$ as the training set. The remaining 20 images serve as the test set.

In addition, we extend the above three public datasets with corresponding text data to provide rich semantic information. We align text prompts with MSI images using CLIP, where Section of the paper details the Adapter fine-tuning method employed. Additionally, we augmented the training set through rotation and flipping.

Implementation details

Following the general setting, we utilize the Nikon D700 to generate HR-MSI for the CAVE and Harvard datasets, and the Landsat-8 spectral response function to generate HR-MSI for the Chikusei dataset. During training, we augment the data with random horizontal, vertical flips and 90-degree rotations. The Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is adopted. The initial learning rate is set to 1×10^{-4}. In addition, the model is trained on a NVIDIA GeForce A100 GPU using the PyTorch framework.

Comparison with SOTA Methods

To verify the superior performance of our proposed CO^{2}IF, experiments are conducted on the CAVE, Chikusei, and Harvard datasets. We select three traditional numerical methods: GSA (Aiazzi, Baronti, and Selva 2007), CNMF (Yokoya, Yairi, and Iwasaki 2011), ICCV15 (Lanaras, Baltsavias, and Schindler 2015); and five deep learning-based methods: BDT (Deng et al. 2023), SSTF-Unet (Liu et al. 2023), FusionMamba (Peng et al. 2024), S^{2}CycleDiff (Qu et al. 2024) and SRLF (Liu et al. 2025).

Quantitative Comparison. We adopt four picture quality indices for quantitative evaluation (Wang et al. 2004): the *peak signal-to-noise ratio* (PSNR), *structural similarity*

Methods	CAVE				Chikusei				Harvard			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
FUSE	38.46	0.9293	5.179	5.437	32.88	0.9253	6.621	7.572	38.80	0.9424	4.562	5.341
CNMF	40.03	0.9432	4.717	5.545	34.78	0.9377	6.412	7.389	39.75	0.9377	4.360	4.974
ICCV15	40.33	0.9446	4.614	5.135	35.52	0.9315	6.228	7.183	40.24	0.9363	4.124	4.765
BDT	48.16	0.9822	2.096	1.891	42.89	0.9843	2.982	3.589	44.52	0.9547	2.863	2.766
SSTF-Unet	48.57	0.9835	2.295	1.801	42.69	0.9827	3.011	3.991	44.81	0.9567	2.859	2.594
FusionMamba	48.85	0.9859	1.981	1.776	43.53	0.9864	2.837	3.145	45.04	0.9588	2.739	2.429
S^2CycleDiff	48.66	0.9854	1.993	1.798	42.72	0.9838	2.976	3.624	44.95	0.9592	2.868	2.611
SRLF	49.19	0.9864	1.760	1.695	43.08	0.9857	2.856	3.326	45.26	0.9607	2.638	2.306
CO^2IF (Ours)	**50.25**	**0.9886**	**1.436**	**1.462**	**44.76**	**0.9891**	**2.624**	**2.885**	**46.45**	**0.9668**	**2.413**	**2.055**

Table 1: The average quantitative performance of various methods on CAVE, Chikusei and Harvard datasets. The best results is highlighted in bold. A higher PSNR and SSIM indicates better results, and a lower SAM and ERGAS indicates better results.

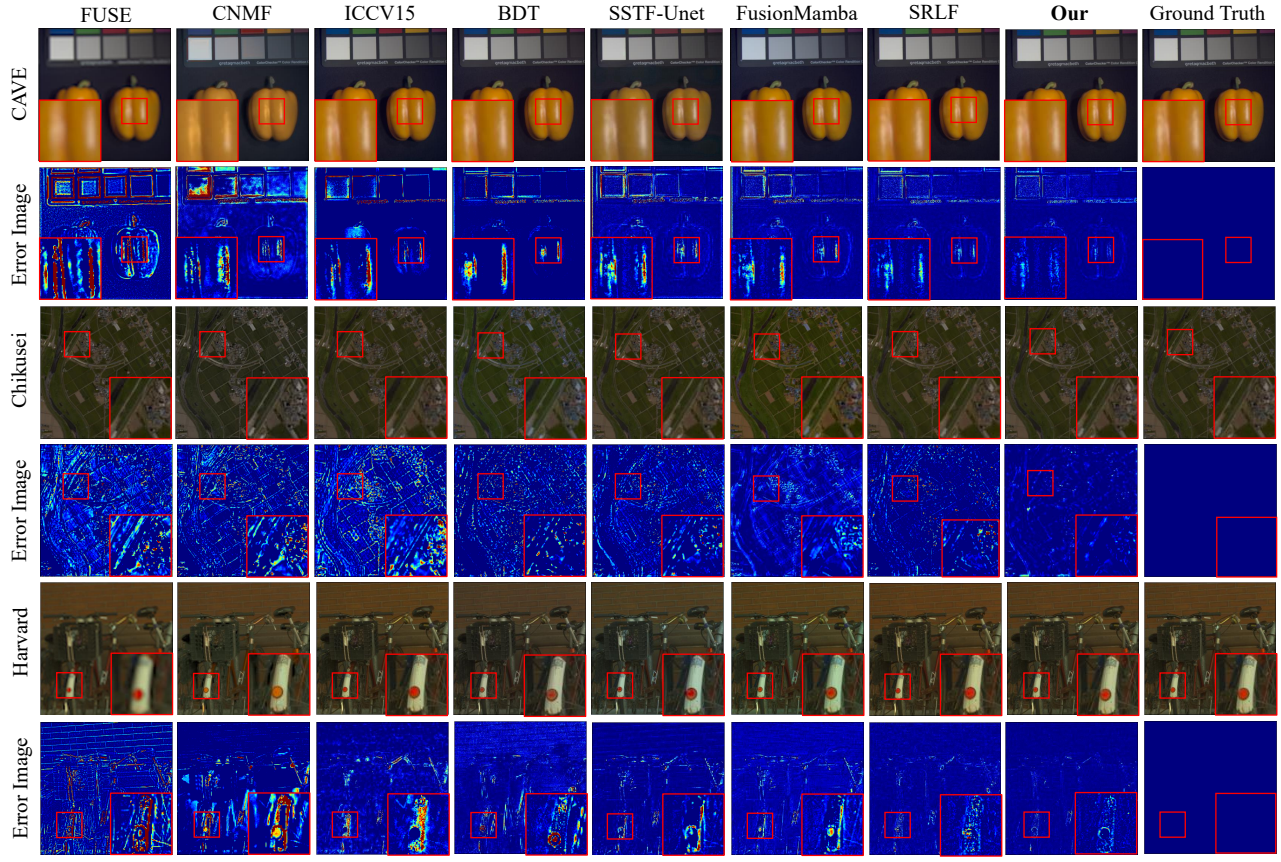

Figure 4: Visualization of HSI reconstruction results and the corresponding mean squared error (MSE) images for various HS/MS fusion methods on the CAVE, Chikusei and Harvard datasets.

index (SSIM), *spectral angle mapper* (SAM), and *relative global synthesis error* (ERGAS). We average the results of all test samples, and the quantitative comparison of different methods is shown in Table 1. It can be seen that our CO^2IF consistently outperforms all competing methods across four evaluation metrics, demonstrating the excellence of incorporating language-derived prior knowledge in HS/MS fusion.

Qualitative Comparison. To evaluate the visual performance of CO^2IF, we visualize example results of different methods in Figure 4. Error maps beneath the visual results demonstrate deviations from the ground truth. While competing methods demonstrate potential in reconstructing the background, deviations in detail preservation are clearly visible in the error images. Additionally, some methods exhibit discrepancies in color compared to the ground truth. In contrast, our proposed CO^2IF maintains impressive spatial and spectral fidelity, validating the effectiveness of language knowledge in understanding hyperspectral scenes.

Method	PSNR↑	SSIM↑	SAM↓	ERGAS↓
w/o COT	48.26	0.9825	2.023	1.841
w/ CE	48.78	0.9860	1.972	1.748
w/o $\mathcal{G}$	49.68	0.9873	1.576	1.605
CO^2IF (*full*)	**50.25**	**0.9886**	**1.436**	**1.462**

Table 2: Effect of COT in the proposed method.

Method	PSNR↑	SSIM↑	SAM↓	ERGAS↓	#Flops(G)
Convolution	48.32	0.9826	2.164	1.854	86.83
Cross-attention	48.81	0.9857	1.978	1.782	124.55
Cobra	49.45	0.9866	1.597	1.679	111.75
w/o $\mathcal{W}$	49.83	0.9876	1.525	1.641	98.48
CO^2IF (*full*)	**50.25**	**0.9886**	**1.436**	**1.462**	106.67

Table 3: Effect of CoMamba in the proposed method.

Ablation Study

Effect of COT. To verify the effectiveness of COT, we develop several variants by removing or replacing its key components: (1) a variant without COT, (2) a variant using cross-entropy loss (*w/* CE) as a replacement, and (3) a variant without the $\mathcal{G}(\cdot,\cdot)$-constrained cost matrix (*w/o* $\mathcal{G}$). As shown in Table 2, the COT method consistently outperforms these variants on the CAVE dataset, demonstrating that COT effectively strengthens the connection between text and HSIs while enhancing the fine-grained details in the generated HR-HSI.

Effect of CoMamba. To verify the effectiveness of CoMamba, we conduct ablation studies exploring alternative integration strategies: (1) replacing CoMamba with convolution, (2) substituting cross-attention, (3) implementing a cross-modal Mamba that concatenates visual-text tokens (e.g., Cobra (Zhao et al. 2024)), and (4) removing the adaptive weight matrix $\mathcal{W}$. As evidenced by Table 3, our method consistently outperforms all alternatives, confirming its superior capability in fusing multimodal representations.

Modality Ablation. We analyze the impact of text and HR-MSI modalities on LR-HSI reconstruction by removing one modality at a time. As shown in Figure 5, removing either the text or HR-MSI modality results in performance degradation, indicating that the text modality provides rich semantic information, serving as spatial visual guidance similar to HR-MSI. When all three modalities coexist, they complement each other to achieve optimal reconstruction, demonstrating the effectiveness of our method.

Effect of Adaptive Semantic Weight Matrix. We visualize salient features and spectral response curves across bands using Chikusei dataset scenes (Figure 6). The adaptive semantic weight matrix $\mathcal{W}$ enhances spectral representation by selectively highlighting high-variation regions and weighting them within corresponding bands. Spectral response curves confirm significant improvements: Our method with $\mathcal{W}$ (Our_W) achieves responses substantially closer to ground truth than both its ablated version (Our) and other methods, demonstrating superior spectral fidelity.

Effect of Hyperparameters in CoMamba. We analyze

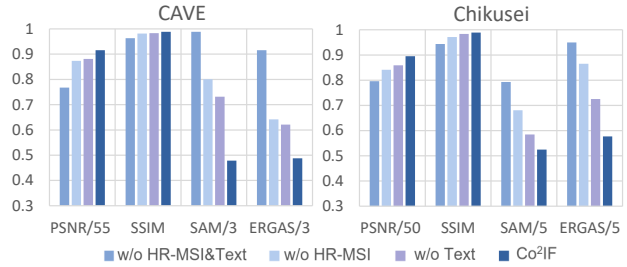

Figure 5: Ablation studys on the influence of HR-MSI and text on the reconstruction results.

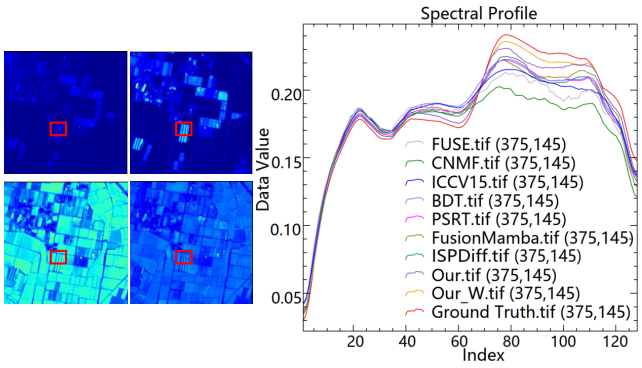

Figure 6: Effect of adaptive semantic weight matrix. The features of selected channels are presented on the left, alongside their spectral response curves in the red box on the right.

λ	0.01	0.1	0.2	γ	0.01	0.1	0.2
PSNR	49.41	**50.25**	49.92	PSNR	49.26	**50.25**	49.87
SSIM	0.9869	**0.9886**	0.9880	SSIM	0.9863	**0.9886**	0.9876

Table 4: Effects of hyperparameters λ and γ in CoMamba.

the values of the hyperparameters λ and γ in the CoMamba state-space formula, as shown in Table 4. The results indicate that the model performs best when both $\lambda = 0.1$ and $\gamma = 0.1$. Therefore, we set 0.1 as the optimal value for both.

Conclusion

In this paper, we propose language-derived hyperspectral and multispectral image fusion in the coordinated state space with cross-modal optimal transport. The proposed COT aligns text with hyperspectral features, assigning semantics to spectral visual cues. We design a CoMamba, guided by text and adaptive semantic weights, to minimize artifacts and achieve high-precision HR-HSI reconstruction. Experiments on public datasets demonstrate the SOTA performance of CO^2IF in both visual quality and objective metrics. In the future, we may explore the application of language in HS/MS fusion tasks using semi-supervised or unsupervised methods. For the datasets, we will aim to enrich a broader range of publicly available datasets by generating and providing corresponding textual descriptions.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFA100860; in part by the National Natural Science Foundation of China under Grant 92470108, Grant 62272363, and Grant 62571395; in part by the Joint Laboratory for Innovation in Satellite-Borne Computers and Electronics Technology Open Fund 2023 under Grant 2024KFKT001-1.

References

Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; and Selva, M. 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5): 591–596.

Aiazzi, B.; Baronti, S.; and Selva, M. 2007. Improving component substitution pansharpening through multivariate regression of MS + Pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3230–3239.

Akhtar, N.; Shafait, F.; and Mian, A. 2015. Bayesian sparse representation for hyperspectral image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3631–3640.

Bashir, A. K.; Victor, N.; Bhattacharya, S.; Huynh-The, T.; Chengoden, R.; Yenduri, G.; Maddikunta, P. K. R.; Pham, Q.-V.; Gadekallu, T. R.; and Liyanage, M. 2023. Federated learning for the healthcare metaverse: Concepts, applications, challenges, and future directions. *IEEE Internet of Things Journal*, 10(24): 21873–21891.

Chakrabarti, A.; and Zickler, T. 2011. Statistics of real-world hyperspectral images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 193–200. IEEE.

Chen, Z.; Pu, H.; Wang, B.; and Jiang, G.-M. 2014. Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods. *IEEE Geoscience and Remote Sensing Letters*, 11(8): 1418–1422.

Deng, S.; Deng, L.-J.; Wu, X.; Ran, R.; and Wen, R. 2023. Bidirectional Dilation Transformer for Multispectral and Hyperspectral Image Fusion. In *IJCAI*, 3633–3641.

Dian, R.; Li, S.; Fang, L.; and Wei, Q. 2019. Multispectral and hyperspectral image fusion with spatial-spectral sparse representation. *Information Fusion*, 49: 262–270.

Dong, W.; Fu, F.; Shi, G.; Cao, X.; Wu, J.; Li, G.; and Li, X. 2016. Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transactions on Image Processing*, 25(5): 2337–2352.

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

Guo, H.; Guo, Y.; Zha, Y.; Zhang, Y.; Li, W.; Dai, T.; Xia, S.-T.; and Li, Y. 2024. MambaIRv2: Attentive State Space Restoration. *arXiv preprint arXiv:2411.15269*.

Han, X.; Yu, J.; Xue, J.-H.; and Sun, W. 2020. Hyperspectral and multispectral image fusion using optimized twin dictionaries. *IEEE Transactions on Image Processing*, 29: 4709–4720.

Hu, B.; Liu, H.; Zheng, Z.; and Liu, P. 2024a. CLIP-SR: Collaborative Linguistic and Image Processing for Super-Resolution. *arXiv preprint arXiv:2412.11609*.

Hu, B.; Liu, H.; Zheng, Z.; and Liu, P. 2024b. CLIP-SR: Collaborative Linguistic and Image Processing for Super-Resolution. *arXiv preprint arXiv:2412.11609*.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6): 066138.

Lanaras, C.; Baltsavias, E.; and Schindler, K. 2015. Hyperspectral super-resolution by coupled spectral unmixing. In *Proceedings of the IEEE international conference on computer vision*, 3586–3594.

Liu, H.; Feng, C.; Dian, R.; and Li, S. 2023. SSTF-Unet: Spatial–spectral transformer-based U-Net for high-resolution hyperspectral image acquisition. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, Y.; Liu, J.; Dian, R.; and Li, S. 2025. A Selective Relearning Mechanism for Hyperspectral Fusion Imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7437–7446.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.

Loncan, L.; De Almeida, L. B.; Bioucas-Dias, J. M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Fabre, S.; Liao, W.; Licciardi, G. A.; Simoes, M.; et al. 2015. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3): 27–46.

Peng, S.; Guo, C.; Wu, X.; and Deng, L.-J. 2023. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3219–3227.

Peng, S.; Zhu, X.; Deng, H.; Deng, L.-J.; and Lei, Z. 2024. FusionMamba: Efficient Remote Sensing Image Fusion With State Space Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Qiao, Y.; Yu, Z.; Guo, L.; Chen, S.; Zhao, Z.; Sun, M.; Wu, Q.; and Liu, J. 2024. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*.

Qu, J.; He, J.; Dong, W.; and Zhao, J. 2024. S2CycleDiff: Spatial-Spectral-Bilateral Cycle-Diffusion Framework for Hyperspectral Image Super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4623–4631.

Qu, Y.; Qi, H.; and Kwan, C. 2018. Unsupervised sparse dirichlet-net for hyperspectral image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2511–2520.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *IJCAI*, volume 1, 2.

Tang, L.; Yin, Z.; Su, H.; Lyu, W.; and Luo, B. 2024. Wfss: weighted fusion of spectral transformer and spatial self-attention for robust hyperspectral image classification against adversarial attacks. *Visual Intelligence*, 2(1): 5.

Wang, G.; Badal, A.; Jia, X.; Maltz, J. S.; Mueller, K.; Myers, K. J.; Niu, C.; Vannier, M.; Yan, P.; Yu, Z.; et al. 2022. Development of metaverse for intelligent healthcare. *Nature machine intelligence*, 4(11): 922–929.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Weng, Z.; Yang, X.; Li, A.; Wu, Z.; and Jiang, Y.-G. 2023. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*, 36978–36989. PMLR.

Wu, C.; Wang, D.; Bai, Y.; Mao, H.; Li, Y.; and Shen, Q. 2023a. HSR-Diff: Hyperspectral image super-resolution via conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7083–7093.

Wu, X.; Zhu, F.; Zhao, R.; and Li, H. 2023b. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7031–7040.

Xiao, J.; and Wei, X. 2023. Hyperspectral Image Denoising Using Uncertainty-Aware Adjustor. In *IJCAI*, 1560–1568.

Xu, R.; Liu, P.; Wang, L.; Chen, C.; and Wang, J. 2020. Reliable Weighted Optimal Transport for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan, H.-F.; Zhao, Y.-Q.; Chan, J. C.-W.; Kong, S. G.; EI-Bendary, N.; and Reda, M. 2025. Hyperspectral and multispectral image fusion: When model-driven meet data-driven strategies. *Information Fusion*, 116: 102803.

Yasuma, F.; Mitsunaga, T.; Iso, D.; and Nayar, S. K. 2010. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9): 2241–2253.

Yokoya, N.; and Iwasaki, A. 2016. Airborne hyperspectral data over Chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5.

Yokoya, N.; Yairi, T.; and Iwasaki, A. 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2): 528–537.

Zhang, M.; Li, L.; Shi, W.; Guo, J.; Li, Y.; and Gao, X. 2024. VmambaSCI: Dynamic Deep Unfolding Network with Mamba for Compressive Spectral Imaging. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6549–6558.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2019a. Deep latent low-rank representation for face sketch synthesis. *IEEE transactions on neural networks and learning systems*, 30(10): 3109–3123.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2019b. Neural probabilistic graphical model for face sketch synthesis. *IEEE transactions on neural networks and learning systems*, 31(7): 2623–2637.

Zhang, M.; Wu, Q.; Guo, J.; Li, Y.; and Gao, X. 2022a. Heat transfer-inspired network for image super-resolution reconstruction. *IEEE Transactions on neural networks and learning systems*, 35(2): 1810–1820.

Zhang, M.; Wu, Q.; Zhang, J.; Gao, X.; Guo, J.; and Tao, D. 2022b. Fluid micelle network for image super-resolution reconstruction. *IEEE Transactions on Cybernetics*, 53(1): 578–591.

Zhang, M.; Xin, J.; Zhang, J.; Tao, D.; and Gao, X. 2022c. Curvature consistent network for microscope chip image super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 10538–10551.

Zhao, H.; Zhang, M.; Zhao, W.; Ding, P.; Huang, S.; and Wang, D. 2024. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16793–16803.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.