

CLIPPan: Adapting CLIP as A Supervisor for Unsupervised Pansharpening

Lihua Jian¹, Jiabo Liu¹, Wushao Wu², Lihui Chen^{3*}

¹School of Electrical and Information Engineering, Zhengzhou University, China

²School of computer science, Wuhan University, China

³School of Microelectronics and Communication Engineering, Chongqing University, China
lihui.chen@cqu.edu.cn

Abstract

Despite remarkable advancements in supervised pansharpening neural networks, these methods face domain adaptation challenges of resolution due to the intrinsic disparity between simulated reduced-resolution training data and real-world full-resolution scenarios. To bridge this gap, we propose an unsupervised pansharpening framework, CLIPPan, that enables model training at full resolution directly by taking CLIP, a visual-language model, as a supervisor. However, directly applying CLIP to supervise pansharpening remains challenging due to its inherent bias toward natural images and limited understanding of pansharpening tasks. Therefore, we first introduce a lightweight fine-tuning pipeline that adapts CLIP to recognize low-resolution multispectral, panchromatic, and high-resolution multispectral images, as well as to understand the pansharpening process. Then, building on the adapted CLIP, we formulate a novel *loss integrating semantic language constraints*, which aligns image-level fusion transitions with protocol-aligned textual prompts (e.g., Wald’s or Khan’s descriptions), thus enabling CLIPPan to use language as a powerful supervisory signal and guide fusion learning without ground truth. Extensive experiments demonstrate that CLIPPan consistently improves spectral and spatial fidelity across various pansharpening backbones on real-world datasets, setting a new state of the art for unsupervised full-resolution pansharpening.

Code— <https://github.com/Jiabo-Liu/CLIPPan>

Introduction

Pansharpening fuses multispectral (MS) images with rich spectral information and panchromatic (PAN) images with detailed spatial information to achieve high-resolution MS (HRMS) images, showing significant potential in remote sensing applications, such as urban planning and environmental surveillance (Vivone et al. 2024).

Over the years, lots of methods have been developed for pansharpening, including component substitution (CS) (Aiazzi, Baronti, and Selva 2007; Kwarteng and Chavez 1989; Carper 1990), multi-resolution analysis (MRA) (Vivone et al. 2015; Otazu et al. 2005; Aiazzi et al. 2006), variational optimization (VO) (Ballester et al. 2006;

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

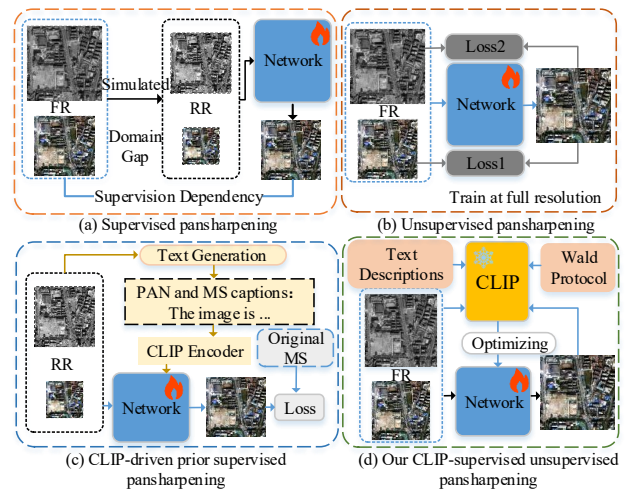


Figure 1: The comparison of the different pansharpening paradigms.

Shutao, Yang, and Bin 2011), and deep learning (DL)-based approaches (Giuseppe et al. 2016; Cao et al. 2025). Nevertheless, the former three types of methods usually struggle to balance spectral and spatial fidelity or rely on hand-crafted priors. In contrast, DL-based methods not only avoid these limitations but also dominate this topic recently for their superior ability to learn complex mappings (Yang et al. 2017).

However, as shown in Figure 1(a), most DL-based methods rely on ground truth (GT) for supervision of the models, such as the classical PNN (Giuseppe et al. 2016), Pan-Net (Yang et al. 2017), and SDRCNN (Fang, Cai, and Fan 2023). Despite the continuous proposal of various methods, most of them focus on architecture designing (Chen et al. 2025) or novel regularization terms (Zeng et al. 2025a) instead of eliminating the dependency on GT. Unfortunately, the GT is inaccessible for MS images in the real scenario at full resolution. To this end, supervised methods are usually trained by simulated data of LRMS, PAN, and HRMS images at reduced resolution, leading to a significant performance degradation on real full-resolution imagery due to domain gaps of scale.

Therefore, unsupervised methods that directly train models at full resolution have attracted the attention of re-

searchers for pansharpening. To avoid the reliance on GT, most unsupervised methods generally train models by employing GAN-based architectures (Ma et al. 2020), priors (Wang et al. 2025a), or spectral-spatial consistency (Ciotola et al. 2022) between the fused and source images. However, due to the absence of the GT, these unsupervised methods (Figure 1 (b)) lack direct guidance to learn pansharpening. Besides, these methods only impose the regularization (Ciotola, Poggi, and Scarpa 2023a; Shen et al. 2024; Barimani and Aghagolzadeh 2025) by a low-level relationship between the fused output and the source images, thus hardly ensuring the output in the HRMS domain.

To address these challenges, a potential solution is to tell the pansharpening model the fusion objective or rule (e.g., Wald’s or Khan’s protocols). In this way, high-level semantic supervision with texture prompts can be utilized to constrain the fused output in the HRMS domain when GT is inaccessible. Therefore, inspired by the remarkable capability of aligning images with texture prompts in a shared semantic space of recent vision-language models, particularly CLIP, we explore an unsupervised pansharpening framework at full resolution by providing textual prompts (e.g., Wald’s protocol) as supervisory signals for pansharpening models in the absence of GT.

However, despite its superior alignment capability, CLIP’s strong bias towards natural images and lack of understanding for pansharpening hinder it a competent supervisor to guide the pansharpening model. Moreover, trained by RGB images, CLIP cannot tackle MS images with more bands and recognize the spectral characters of MS images. Therefore, we first adapt the pre-trained CLIP model to a qualified supervisor by *i)* establishing reliable modality recognition by binding LRMS, PAN, and HRMS images to their corresponding semantic spaces represented by textual prompts through inter-modal contrastive learning (Inter-MCL); *ii)* enabling CLIP to recognize image content of remote sensing images and maintain feature diversity via intra-modal contrastive learning (IntraMCL); and *iii)* guiding CLIP to understand the pansharpening by aligning fused image features with texture prompts of fusion protocols such as Wald’s rule. Once adapted, CLIP serves as a fixed semantic supervisor to guide the pansharpening network via a joint loss integrating low-level visual and semantic texture constraints, effectively bridging the fused outputs with the HRMS domain without requiring any GT labels.

In conclusion, the contributions are as follows.

- We present CLIPPan, a universal framework to leverage vision-language models (particularly CLIP) for unsupervised full-resolution pansharpening via protocol-informed linguistic guidance. Compatible with any pansharpening backbone, CLIPPan achieves state-of-the-art performance on various datasets, significantly enhancing both spectral and spatial fidelity in real-world scenarios.
- We design a lightweight CLIP adaptation strategy tailored for pansharpening, enhancing CLIP’s ability to recognize remote sensing images and the pansharpening process, paving the way for future research in unsupervised pansharpening based on visual-language models.

- We introduce a novel language-guided unsupervised loss based on Wald’s protocols, which provides semantic alignment between the fused output and the HRMS domain, contributing to future works in utilizing textual prompts to supervise the pansharpening. Besides, this research can establish a reciprocal framework. That is, the underlying paradigm can conversely evaluate the effectiveness of protocols and even guide the discovery of novel pansharpening protocols.

Related Works

DL-Based Pansharpening

Optimized by pixel-level losses (e.g., ℓ_1 or ℓ_2) between the fused result and the ground truth, supervised methods are trained at reduced resolution using paired triplets of LRMS, PAN, and HRMS images. Representative works include PNN (Giuseppe et al. 2016), PanNet (Yang et al. 2017), and SDRCNN (Fang, Cai, and Fan 2023) as well as more recent architectures like ADKNet (Peng et al. 2022), PSCINN (Wang et al. 2024), SSFMamba (Ma et al. 2025), and FusionMamba (Peng et al. 2024). Despite strong performance on synthetic datasets, these methods suffer from poor generalization to real full-resolution images due to the domain gap between the simulated reduced-resolution and real full-resolution data. Although a recent study (Cao et al. 2025) shown in Figure 1 (c) attempted to introduce CLIP-derived priors, it still dependent on low-resolution settings and handcrafted protocols, leaving the full-resolution unsupervised setting underexplored. Differently, our CLIPPan shown in Figure 1 (d) allows direct training at full resolution with both semantic and low-level visual constraints, eliminating reliance on GT while improving spectral-spatial fidelity.

Unsupervised methods avoid using HRMS labels and instead design training objectives based on low-level relationships between the fused image and its inputs, such as spatial-spectral consistency (Ciotola et al. 2022), modality alignment (Zeng et al. 2025b), or detail preservation (Lin et al. 2024). Some approaches adopt handcrafted priors (Zeng et al. 2025a), while others use generative frameworks like GANs (Ma et al. 2020; Zhu et al. 2023; Liu et al. 2025) or diffusion models (Wang et al. 2025b; Zhang et al. 2025) to better model fusion distributions. However, these methods still rely on heuristic, low-level visual constraints that offer limited guidance on what constitutes a semantically valid or perceptually high-quality fusion.

Visual-Language Alignment

While vision-language models like CLIP (Radford et al. 2021) have demonstrated powerful generalization across modalities and domains, directly applying them to remote-sensing tasks remains non-trivial. This is due to both the modality gap—satellite imagery significantly differs from natural images in texture, geometry, and semantics—and the task gap, as standard CLIP is trained for image-text alignment rather than pansharpening.

To adapt CLIP for downstream tasks, numerous parameter-efficient fine-tuning strategies have been proposed. CLIP-Adapter (Gao et al. 2024) inserts lightweight

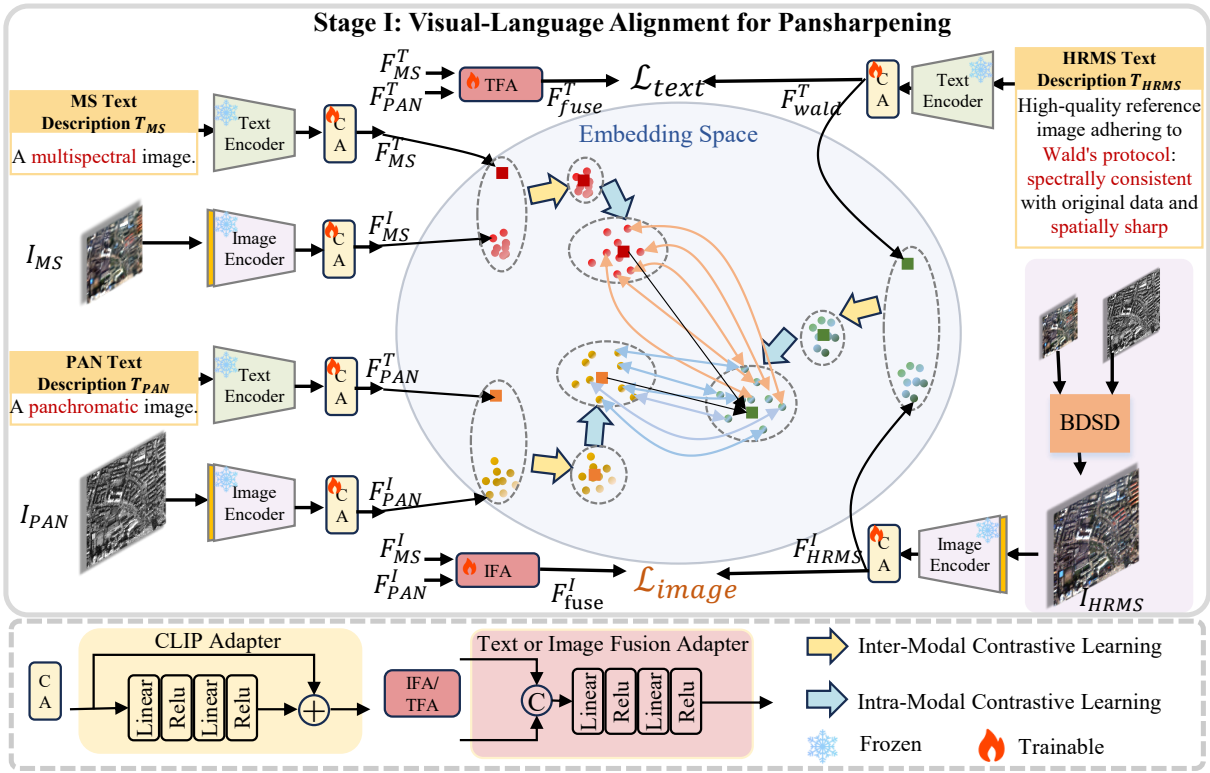


Figure 2: Workflow for stage I. Visual-language alignment for pansharpening.

bottleneck layers to inject task-specific knowledge; prompt-based tuning methods such as CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) learn input-dependent textual prompts to steer the joint embedding space; LoRA-CLIP (Zanella and Ben Ayed 2024) applies low-rank adaptation to reduce training overhead. In remote sensing, recent methods like RS-CLIP (He et al. 2024) and GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023) incorporate domain-specific prompts or geospatial priors to bridge the semantic gap between satellite images and natural images.

Unlike prior works that focus on classification or segmentation, we take a novel step by adapting CLIP as a semantic supervisor for an unsupervised pansharpening task.

Methodology

As shown in Figure 2 and Figure 3, the proposed CLIP-Pan achieves unsupervised pansharpening by two stages. At the first stage, CLIP is adapted to align the LRMS, PAN, and HRMS images with the corresponding semantic space described by texture prompts, respectively. Notably, the HRMS images are particularly aligned to a semantic space described by the pansharpening objective, e.g., Wald’s protocol. Then, at the second stage, the adapted CLIP supervises the pansharpening model by evaluating whether the fused out align with the pansharpening objective or not, thus eliminating the dependence on the GT.

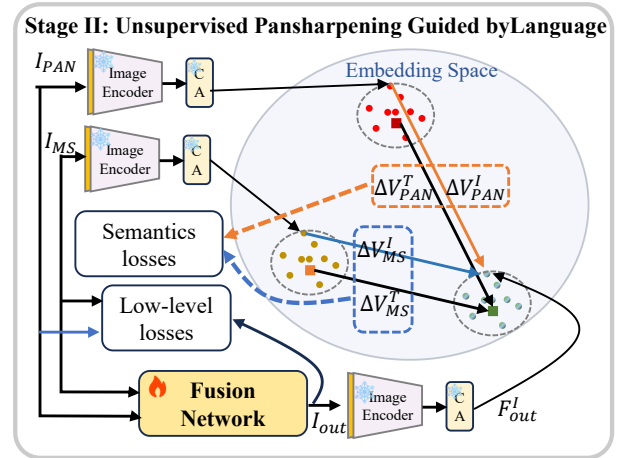


Figure 3: Workflow for stage II. Unsupervised pansharpening guided by language.

Stage I: Visual-Language Alignment for Pansharpening

Despite superior alignment performance on web images and language, the pretrained CLIP face challenges of bias toward natural images and incompatibility with MS image bands for the purpose. Therefore, adapting CLIP is critical for pansharpening. Meanwhile, to maintain the strong generalization of CLIP, a parameter-efficient fine-tuning strategy

is adopted in our adaptation. Specifically, we introduce six lightweight adapter modules (i.e., CA modules in Figure 2): three after the visual encoder for vision adaptation, and the other three after the text encoder for text adaptation. Besides, since the visual encoder is incompatible with the MS images, we replace the original layer designed for natural RGB images with a convolutional layer specific for MS inputs.

To make the CLIP a proficient supervisor for pansharpening, the adaptation of is optimized by the following loss:

$$\mathcal{L}_{s1} = \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{fusion}}, (1)$$

where $\mathcal{L}_{\text{intra}}$, the InterMCL loss, is used to bind image types of LRMS, PAN, and HRMS to their semantic spaces, respectively; $\mathcal{L}_{\text{inter}}$, the IntraMCL loss, is used to ensure the adapted CLIP to recognize image content and feature diversity; and $\mathcal{L}_{\text{fusion}}$, the fusion loss, is employed to ensure that the feature representations of MS and PAN images can be projected into the HRMS feature space.

Inter-Modal Contrastive Learning (InterMCL). To adapt CLIP to remote sensing images, an intuitive attempt is to utilize the vanilla language-image contrastive learning in CLIP. However, for finetuning a pansharpening supervisor, the vanilla way is inappropriate to adapt CLIP. Since, as a pansharpening supervisor, the model needs the ability to discriminate the MS, PAN, and HRMS images, thereby utilizing its understanding of image types to supervise the learning of pansharpening models. Nevertheless, the vanilla contrastive language-image learning utilizes content-dependent descriptions for various images, aiming at recognizing image content by language prompts, which is inconsistent with the objective of recognizing different image types. Besides, it is difficult to create a dataset with quadruples of LRMS-PAN-HRMS-text.

Therefore, to bind an image type to a semantic space, a better choice is to use semantically similar descriptions for a distinct image type instead of content-dependent descriptions for various images. For simple, constant descriptions for distinct types of images are used in the proposed method texture, i.e., $F_{\text{MS}} = \text{“}a \text{ multispectral image”}$, $T_{\text{PAN}} = \text{“}a \text{ panchromatic image”}$, $T_{\text{HRMS}} = \text{“}High\text{-quality reference image adhering to Wald’s protocol: spectrally consistent with original data and spatially sharp”}$. Meanwhile, considering that HRMS images cannot be obtained at full resolution in real cases, HRMS images are generated on-the-fly using the conventional BDSD (Vivone 2019) algorithm.

Specifically, as shown in Figure 2, we pair every image in a specific type, e.g., F_{MS}^I , with its corresponding text prompt (i.e., F_{MS}^T). Then, these matched pairs are treated as positives, while all other image-text combinations within the batch serve as negatives. This encourages the model to pull together the paired image-text samples semantically aligned representations while pushing apart mismatched ones, thus enabling CLIP to recognize image types and binding image types to the corresponding semantic spaces. In detail, the

inter-MCL can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{inter}} &= \frac{1}{3} \sum_{M1, M2} \mathcal{L}_{\text{align}}(F_{M1}^I, F_{M2}^T), \\ \mathcal{L}_{\text{align}} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\langle F_{M1}^{I(i)}, F_{M2}^{T(i)} \rangle / \tau_c)}{\sum_{j=1}^N \exp(\langle F_{M1}^{I(i)}, F_{M2}^{T(j)} \rangle / \tau_c)}, \end{aligned} (2)$$

where $M1, M2 \in \{\text{MS, PAN, HRMS}\}$; $F_{M1}^{I(i)}$ and $F_{M2}^{T(i)}$ are the i -th adapted $M1$ -type image and $M2$ -type text features, $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b} / (\|\mathbf{a}\| \|\mathbf{b}\|)$ computes the cosine similarity between vectors, and τ is the temperature controlling alignment strength. This loss ensures that the adapted CLIP preserves language-image alignment necessary for downstream semantic supervision.

Intra-Modal Contrastive Learning (IntraMCL). However, using constant descriptions for all images of a specific image type to adapt CLIP easily results in feature collapsing to the fixed embedding representation of its corresponding texture prompt. Besides, it decreases CLIP’s capability of recognizing image content.

Therefore, we additionally introduce contrastive learning within the image domain by taking the same scene of LRMS, PAN, and HRMS images as positive samples, while others as negative samples. Therefore, images with similar geographic scenes are closing while images with different scenes are diverging, thus ensuring diversity of image features and discriminating different image semantic content. Besides, IntraMCL facilitates domain transfer from natural imagery to remote sensing and bridges the gap from natural to remote sensing domain.

Specifically, we construct training batches by sampling N paired triplets consisting of LRMS, PAN, and HRMS patches. Then, the IntraMCL can be formulated as:

$$\mathcal{L}_{\text{intra}} = -\frac{1}{3N} \sum_{i=1}^{3N} \log \frac{\exp(\langle F_{M1}^{I(i)}, F_{M2}^{I(i)} \rangle / \tau_i)}{\sum_{k=1}^{3N} \exp(\langle F_{M1}^{I(i)}, F_{M1}^{I(j)} \rangle / \tau_i)}. (3)$$

Fusion-Aware Alignment. Although InterMCL and IntraMCL bind various image types to a specific semantic space while maintaining feature diversity, it does not explicitly model the fusion between MS and PAN inputs, lacking understanding of fusion processing. Therefore, we enforce the fusion learning in the adaptation of CLIP, so that CLIP can recognize the fused HRMS representation as a meaningful combination of its sources.

To this end, we introduce two auxiliary modules, i.e., the image fusion adapter (IFA) and the text fusion adapter (TFA) in Figure 2. These adapters operate on encoded features and learn to generate fused image and text embeddings from LRMS and PAN inputs:

$$F_{\text{fuse}}^I = \text{IFA}(F_{\text{MS}}^I, F_{\text{PAN}}^I), F_{\text{fuse}}^T = \text{TFA}(F_{\text{MS}}^T, F_{\text{PAN}}^T), (4)$$

By enforcing alignment between these fused image/text features and the corresponding reference features, i.e., the (HRMS image)/(Wald’s protocol text) features, we guide the model to internalize the mapping from source modalities to

a semantically valid fusion target. In detail, the alignment is achieved through simple \mathcal{L}_1 loss:

$$\mathcal{L}_{\text{fusion}} = \|F_{\text{fuse}}^T - F_{\text{wald}}^T\|_1 + \|F_{\text{fuse}}^I - F_{\text{HRMS}}^I\|_1. (5)$$

Eventually, the adapted CLIP acquires the ability to project MS and PAN image features into the semantic space of high-quality HRMS images, which facilitates better supervision for the pansharpening model in Stage II.

Stage II:

Unsupervised Pansharpening Guided by Language

To achieve unsupervised pansharpening, we combine the semantic and low-level supervision for the pansharpening model and leverage their complementary advantages.

Semantic Supervision by Language. Thanks to the adaptation in Stage I, the fusion objective described in the textual prompt (i.e., Wald’s protocol) is aligned to the domain of HRMS images. Therefore, the adapted CLIP can naturally evaluate the quality of fusion by judging if the fused output is aligned with the semantic features extracted from the Wald’s protocol or not. However, we cannot use element-wise loss between the visual features extracted from the adapted CLIP with features of Wald’s protocol, due to the latter’s invariance for all fused images.

Nevertheless, as show in Figure 3, the vectors, i.e., $\Delta \mathbf{V}_{\text{MS}}^T$, from features of F_{MS}^T to Wald’s F_{wald}^T can reflect the feature transition of fusion. Therefore, we employ a directional vector from source to target text to guide the training of the pansharpening network. Specifically, the pansharpening network can be supervised by minimizing the angular discrepancy between the feature displacement vectors of image pairs and their corresponding text pairs, i.e.

$$\mathcal{L}_d = 1 - \frac{1}{2} (\langle \Delta \mathbf{V}_{\text{MS}}^I, \Delta \mathbf{V}_{\text{MS}}^T \rangle + \langle \Delta \mathbf{V}_{\text{PAN}}^I, \Delta \mathbf{V}_{\text{PAN}}^T \rangle), (6)$$

$$\Delta \mathbf{V}_{\text{MS}}^I = F_{\text{out}}^I - F_{\text{MS}}^I, \Delta \mathbf{V}_{\text{MS}}^T = F_{\text{wald}}^T - F_{\text{MS}}^T, (7)$$

$$\Delta \mathbf{V}_{\text{PAN}}^I = F_{\text{out}}^I - F_{\text{PAN}}^I, \Delta \mathbf{V}_{\text{PAN}}^T = F_{\text{wald}}^T - F_{\text{PAN}}^T. (8)$$

where F_{out}^I denotes the fused image embedding, and F_{MS}^I , F_{PAN}^I be the embeddings of MS and PAN inputs, respectively. F_{wald}^T , F_{MS}^T , and F_{PAN}^T are the textual embeddings, respectively.

Consequently, through penalizing angular misalignment between transitions in the image and text shared embedding spaces, the pansharpening model is encouraged to produce outputs semantically aligned with the HRMS image domain.

Low-Level Unsupervised Reconstruction Losses. However, since the semantic supervision by language can only impose the output in the HRMS image domain, the image content and spectral characters are not supervised. Therefore, we also introduce the low-level visual constraints for unsupervised pansharpening.

Specifically, the low-level visual unsupervised loss function $\mathcal{L}_{\text{unsup}}$ includes three key components to simultaneously enforce spectral fidelity, spatial sharpness, and perceptual quality. The spectral fidelity is achieved by

$$\mathcal{L}_{\text{spec}} = \|\downarrow(\mathbf{I}_{\text{out}}) - \mathbf{I}_{\text{MS}}\|_2^2 + 1 - \text{SSIM}(\downarrow(\mathbf{I}_{\text{out}}), \mathbf{I}_{\text{MS}}), (9)$$

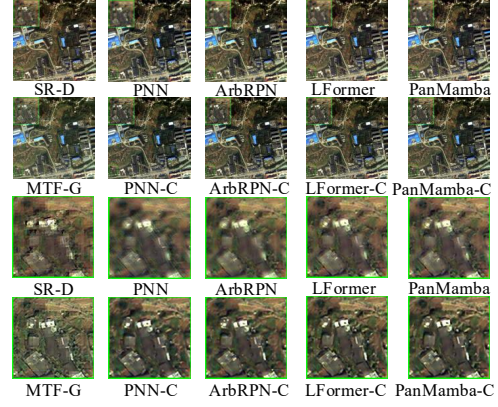


Figure 4: Qualitative results of different methods on the full-resolution QB dataset. The top-row images are shown in RGB, and the areas enclosed by the green box have been magnified three times.

where \downarrow denotes bicubic downsampling (4× ratio), \mathbf{I}_{out} is the fused HRMS output and SSIM (Wang et al. 2004) measures structural similarity at reduced resolution. The spatial sharpness is achieved by:

$$\mathcal{L}_{\text{spat}} = \|\phi(\mathbf{I}_{\text{out}}) - \mathbf{I}_{\text{PAN}}\|_2^2 + 1 - \text{SSIM}(\phi(\mathbf{I}_{\text{out}}), \mathbf{I}_{\text{PAN}}), (10)$$

where $\phi(\cdot)$ denotes a 1×1 convolution that degrades the multispectral channels into a single band. A trade-off loss between spectral and spatial information based on QNR (Alparone et al. 2008) is also adopted

$$\mathcal{L}_{\text{QNR}} = (1 - D_\lambda)(1 - D_s). (11)$$

Finally, to stabilize the training process, we introduce a pseudo-supervision $\mathcal{L}_{\text{ship}}$ that takes as the reference the output of an existing pansharpening network [SHIP (Zhou et al. 2025) used in this paper] trained at reduced resolution. Therefore, the overall low-level visual loss is

$$\mathcal{L}_{s2} = \mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{spat}} + \mathcal{L}_{\text{QNR}} + \mathcal{L}_{\text{ship}}. (12)$$

Experiments

Experimental Settings

Datasets. Our experimental datasets are captured from the WorldView-3 (WV3) sensor, comprising eight spectral bands, as well as a four-band dataset from the QuickBird (QB) sensor. The QB and WV3 datasets are divided into non-overlapped training, validation, and test sets, respectively.

Training Details. All models in paper were trained on a desktop computer equipped with a GTX-4090 GPU. The update of the CLIPPan framework is optimized by the Adam optimizer with a 0.003 learning rate. The batch size and the iteration number are set to 32 and 1000, respectively.

Metrics. For reduced-resolution experiment, we adopt four indicators for assessment: the mean peak signal-to-noise ratio (MPSNR), the erreur relative globale adimensionnelle de synthèse (ERGAS) (Wald 2002), the spectral

Methods	QB								WV3							
	$D_\lambda \downarrow$	$D_s \downarrow$	QNR \uparrow	MPSNR \uparrow	ERGAS \downarrow	SAM \downarrow	Q2n \uparrow		$D_\lambda \downarrow$	$D_s \downarrow$	QNR \uparrow	MPSNR \uparrow	ERGAS \downarrow	SAM \downarrow	Q2n \uparrow	
SR-D	0.0151	0.0410	0.9445	39.0381	3.1441	3.2672	0.7488		0.0258	0.0580	0.9177	29.3070	8.9904	9.6229	0.7334	
MTF-G	0.0386	0.0557	0.9104	44.3701	1.8564	1.2384	0.8088		0.0585	0.0552	0.8895	30.5174	6.2900	6.4423	0.7710	
λ -PNN	0.0604	0.1003	0.8467	42.1526	1.7170	1.5809	0.7263		0.0722	0.0565	0.8757	31.1314	6.5617	6.2032	0.7006	
λ -PNN-C	0.0063	0.0294	0.9643	44.3535	1.3345	1.6468	0.7858		0.0091	0.0364	0.9547	33.9501	4.8606	5.9487	0.7914	
PNN	0.0137	0.0347	0.9521	48.8494	0.7372	0.9404	0.8799		0.0245	0.0358	0.9405	37.1830	3.2368	4.3787	0.8324	
PNN-C	0.0138	0.0336	0.9532	48.9031	0.7226	0.9275	0.8806		0.0098	0.0404	0.9501	37.2778	3.2802	4.3172	0.8355	
ArbRPN	0.0140	0.0281	0.9582	51.2758	0.5470	0.7299	0.9079		0.0271	0.0356	0.9383	38.3459	2.8467	3.7834	0.8538	
ArbRPN-C	0.0030	0.0279	0.9691	51.3047	0.5448	0.7250	0.9083		0.0042	0.0375	0.9582	38.6344	2.7254	3.6221	0.8558	
LFormer	0.0124	0.0277	0.9602	50.2647	0.6210	0.8092	0.8988		0.0253	0.0541	0.9227	38.0189	2.9513	3.8192	0.8493	
LFormer-C	0.0053	0.0272	0.9676	50.3799	0.6149	0.7990	0.9006		0.0049	0.0380	0.9572	38.2575	2.8623	3.7609	0.8506	
PanMamba	0.0134	0.0277	0.9592	50.5325	0.5927	0.7778	0.8998		0.0152	0.0429	0.9426	38.1740	2.8900	3.7766	0.8504	
PanMamba-C	0.0050	0.0278	0.9672	50.6724	0.5888	0.7731	0.9022		0.0051	0.0371	0.9578	38.2407	2.8548	3.7540	0.8513	

Table 1: Quantitative results on the full-resolution and reduced-resolution datasets.

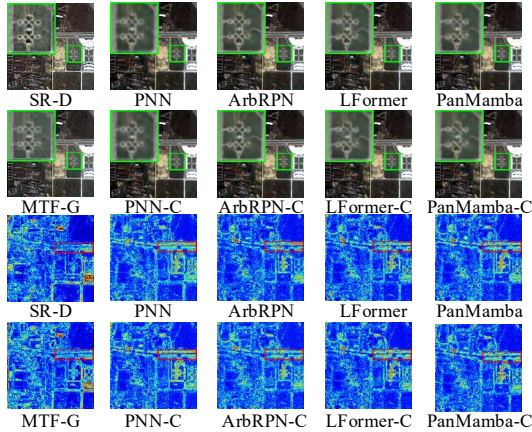


Figure 5: Qualitative results of different methods on the reduced-resolution QB dataset. The top-row images are shown in RGB, and the areas enclosed by the green box have been magnified three times. The bottom-row images are error maps.

angle mapper (SAM) (Yuhua, Goetz, and Boardman 1992), and the Q2n index (Garzelli and Nencini 2009), i.e., Q4 for four band and Q8 for eight band datasets. For full-resolution experiments, we use the quality without a reference (QNR) index (Alparone et al. 2008) and the related spatial (D_s) and spectral (D_λ) distortion indexes.

Effectiveness for Different Baseline Methods

Quantitative Comparison. To validate the universality of the proposed CLIPPan framework, four representative pansharpening backbones— λ -PNN (Ciotola, Poggi, and Scarpa 2023b), PNN (Giuseppe et al. 2016), ArbRPN (Chen et al. 2022), LFormer (Hou et al. 2024), and Pan-Mamba (He et al. 2025) trained with CLIPPan are denoted “Method-C”, while the baseline version employs the conventional loss. For valid and fair comparison, full-resolution and reduced-resolution experiments adopted unsupervised and supervised training approaches, respectively. The traditional SR-D (Vicinanza et al. 2014) and MTF-GLP-HPM-R (Vivone, Restaino, and Chanussot 2018) are employed solely as com-

parative methods. Table 1 reports the quantitative results. Across the board, integrating CLIPPan yields consistent improvements for all backbones. Specifically, compared with baseline methods, ArbRPN-C achieves a 79% reduction in spectral distortion D_λ and enhances QNR by 0.011 on the QB dataset. Similarly, LFormer-C decreases spatial distortion D_s by approximately 30% and improves QNR by 0.035 on the WV3 dataset. These results conclusively demonstrate that the CLIPPan framework has the capability of transferring the language descriptions in Wald’s protocol to pansharpening outcomes. That is to say, even without ground-truth data, the proposed CLIPPan framework still improves spectral and spatial fidelity for the pansharpening task.

Additionally, all evaluation metrics of reduced-resolution experimental results have been improved compared with baseline methods. It shows that the CLIPPan framework is also effective for supervised pansharpening.

Qualitative Comparison. Figure 4 shows the qualitative results on the full-resolution QB dataset. Obviously, results generated by the CLIPPan-integrated (“Method-C”) methods have more spatial detail information than those produced by baseline methods. The magnified regions in the bottom row further reveal that integrating the CLIPPan framework into the baseline methods enhances texture details, sharpens outlines, and improves contrast.

To demonstrate the generalization ability of the proposed CLIPPan framework, we conduct experiments on reduced-resolution QB dataset. As shown in Figure 5, it is observed that the Method-C group consistently exhibits the closest texture details and spectral fidelity to the reference image.

Ablation Study

Effects of Different Unsupervised Loss. We explore how loss in Stage II affects the pansharpening outcomes by progressively adding loss in Eqs. (12) and (6), where the quantitative results are presented in Table 2. The $\mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{spat}}$ combining pixel-wise spectral and spatial reconstruction terms is adopted as the baseline to ensure the spectral fidelity and spatial sharpness of pansharpened results. By integrating the QNR-based loss with these spectral and spatial terms, $\mathcal{L}_{\text{unsup}}$ leads to a significant increase in MPSNR and Q2n by

Method	MPSNR↑	ERGAS↓	SAM↓	Q2n↑
$\mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{spat}}$	29.2739	8.9584	9.1671	0.6102
\mathcal{L}_{QNR}	32.0266	5.9536	6.7052	0.7013
$\mathcal{L}_{\text{unsup}}$	32.1946	5.8776	6.6616	0.7100
$\mathcal{L}_{\text{unsup}} + \mathcal{L}_{\text{ship}}$	33.5735	4.5659	5.9493	0.7661
$\mathcal{L}_{\text{unsup}} + \mathcal{L}_d$	32.3696	5.7517	6.5471	0.7366
$\mathcal{L}_{\text{unsup}} + \mathcal{L}_{\text{ship}} + \mathcal{L}_d$	34.7191	4.4922	5.5429	0.7986

Table 2: Ablation experiment on unsupervised fusion losses. $\mathcal{L}_{\text{unsup}}$ denotes $\mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{spat}} + \mathcal{L}_{\text{QNR}}$.

Method	MPSNR↑	ERGAS↓	SAM↓	Q2n↑
w/o fine-tuning	34.3445	4.7019	5.5517	0.7848
$\mathcal{L}_{\text{intra}}$	34.5553	4.5237	5.5141	0.7904
$\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$	34.6019	4.5349	5.2694	0.7930
$\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_1$	34.7191	4.4922	5.5429	0.7986

Table 3: Ablation experiments on WV3 for adapting CLIP.

2.92dB and 0.10, respectively. A significant reduction is also achieved in ERGAS and SAM, 3.08 and 2.51, respectively. Similarly, separately adding either pseudo-supervision loss $\mathcal{L}_{\text{ship}}$ or language-guided semantic loss \mathcal{L}_d to the unsupervised loss leads to improvements in all metrics. By contrast, incorporating both losses simultaneously into the unsupervised loss significantly improves all metrics. These results confirm that the combination of unsupervised reconstruction, pseudo-label regularization, and semantic alignment via language in our training loss is the best choice.

Effects of CLIP Fine-Tuning Loss. We verify the effectiveness of each loss term by adding them gradually to the CLIP adaptation stage. Table 3 summarizes the results of the ablation experiment on the WV3 dataset. We use the CLIP model without any fine-tuning as the baseline. Therefore, incorporating IntraMCL loss $\mathcal{L}_{\text{intra}}$ improves the model performance across all metrics. These indicate that $\mathcal{L}_{\text{intra}}$ loss promotes the CLIP model to recognize remote-sensing image content, providing valuable semantic information for the downstream pansharpening task. Similarly, incrementally adding InterMCL loss $\mathcal{L}_{\text{inter}}$ further enhances model performance, particularly improving MPSNR by 0.26dB and reducing SAM by 0.28. These demonstrate that $\mathcal{L}_{\text{inter}}$ loss enables fine-grained alignment between visual and textual representations, thereby achieving modality-invariant feature learning. Further applying the \mathcal{L}_1 loss, we observe sustained improvements in all evaluation metrics. These results reflect that the incorporation of IntraMCL, InterMCL, and \mathcal{L}_1 constraints on CLIP fine-tuning achieves best pansharpening performance.

Effects of Different MS Input Manner. To ensure compatibility with CLIP’s three-channel input, we evaluated four strategies for compressing the multi-band MS image, as summarized in Table 4. Results indicate that directly utilizing RGB or GBNIR channels for extraction struggles to achieve satisfactory pansharpening performance. Although

Method	MPSNR↑	ERGAS↓	SAM↓	Q2n↑
PCA	34.6903	4.7911	5.6647	0.7955
RGB	34.4237	4.6220	5.6168	0.7940
GBNIR	34.3282	4.5038	5.6804	0.7957
Conv	34.7191	4.4922	5.5429	0.7986

Table 4: Ablation experiment on feature extraction of MS images in CLIP. PCA: principal component analysis; RGB: direct RGB extraction; GBNIR: Green-Blue-NIR composite; Conv: learnable residual convolution.

Method	MPSNR↑	ERGAS↓	SAM↓	Q2n↑
Noise	34.3671	4.6553	5.8676	0.7869
I	34.5060	4.6490	5.4767	0.7911
II	34.7662	4.5178	5.6075	0.7968
Khan’s	34.6581	4.5021	5.5506	0.7969
Wald’s	34.7191	4.4922	5.5429	0.7986

Table 5: Ablation experiment on textual descriptors for HRMS images. “Noise descriptor” states “an image independent of the inputs,” “I” denotes “This image is the fusion image of the input image,” and “II” denotes “a fused product of the MS and PAN images”.

the PCA approach improves the MPSNR metric, it leads to significant degradation in other metrics, particularly with a substantial drop in ERGAS. In contrast, the proposed learnable residual convolution yields the best trade-off across all metrics, confirming that a data-driven channel projection preserves both spectral fidelity and spatial detail.

Effects of Different Textual Fusion Descriptors. Table 5 lists results of textual fusion descriptions on final performance. The Noise descriptor and Khan’s protocol (Mangolini, Ranchin, and Wald 1995) fail to achieve the best effect on any metric. Unfortunately, neither “I” nor “II” description approach can effectively optimize both MPSNR and SAM metrics simultaneously. Overall, Wald’s method achieved optimal balance in all metrics. These findings demonstrate that precise, protocol-compliant text supervision is crucial for leveraging CLIP’s semantic space in unsupervised pansharpening.

Conclusion

In this paper, we presented CLIPPan, a novel unsupervised pansharpening framework that repurposes CLIP as a language-driven supervisor. By lightweight fine-tuning, CLIP learns to recognize LRMS, PAN, and HRMS types and enforces fusion rules such as Wald’s protocol purely through text. Extensive experiments on real full-resolution imagery show that CLIPPan consistently boosts existing backbones, outperform the SOTA unsupervised framework, and sets new state-of-the-art results without any GT.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62101502 and 62301093, the China Postdoctoral Science Foundation (CPSF) under Grant 2022T150596 and 2023M730425, the Postdoctoral Fellowship Program of CPSF under Grant GZC20233336, the Key Research & Development and Promotion Foundation of He'nan Province under Grant 232102211036, and the Natural Science Foundation of Chongqing under Grant CSTB2024NSCQ-MSX0918, and in part by the Natural Science Foundation of Henan under Grant 252300421876.

References

- Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; and Selva, M. 2006. MTF-tailored Multiscale Fusion of High-resolution MS and Pan Imagery. *Photogramm. Eng. Remote Sens.*, 72: 591–596.
- Aiazzi, B.; Baronti, S.; and Selva, M. 2007. Improving Component Substitution Pansharpening Through Multivariate Regression of MS+Pan Data. *IEEE Trans. Geosci. Remote Sens.*, 45(10): 3230–3239.
- Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; and Selva, M. 2008. Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.*, 74(2): 193–200.
- Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; and Rougé, B. 2006. A Variational Model for P+XS Image Fusion. *Int. J. Comput. Vis.*, 69(1): 43–58.
- Barimani, M.; and Aghagolzadeh, A. 2025. Unsupervised CNN-based pan-sharpening with generative multiadversarial networks: a colorization approach for panchromatic images. *Int. J. Image Data Fusion*, 16(1): 1–26.
- Cao, Z.-H.; Liang, Y.-J.; Deng, L.-J.; and Vivone, G. 2025. An Efficient Image Fusion Network Exploiting Unifying Language and Mask Guidance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–18.
- Carper, W. R. 1990. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogramm. Eng. Remote Sens.*, 56: 457–467.
- Chen, L.; Lai, Z.; Vivone, G.; Jeon, G.; Chanussot, J.; and Yang, X. 2022. ArbRPN: A Bidirectional Recurrent Pansharpening Network for Multispectral Images With Arbitrary Numbers of Bands. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–18.
- Chen, L.; Song, T.; Jian, L.; Zhang, D.; Vivone, G.; and Zhou, X. 2025. High-Fidelity Pansharpening via Trigeminal Pyramid Decoding of CNN-Transformer Encoded Features. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–16.
- Ciotola, M.; Poggi, G.; and Scarpa, G. 2023a. Unsupervised Deep Learning-Based Pansharpening With Jointly Enhanced Spectral and Spatial Fidelity. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–17.
- Ciotola, M.; Poggi, G.; and Scarpa, G. 2023b. Unsupervised Deep Learning-Based Pansharpening With Jointly Enhanced Spectral and Spatial Fidelity. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–17.
- Ciotola, M.; Vitale, S.; Mazza, A.; Poggi, G.; and Scarpa, G. 2022. Pansharpening by Convolutional Neural Networks in the Full Resolution Framework. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–17.
- Fang, Y.; Cai, Y.; and Fan, L. 2023. SDRCNN: A single-scale dense residual connected convolutional neural network for pansharpening. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 16: 6325–6338.
- Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2): 581–595.
- Garzelli, A.; and Nencini, F. 2009. Hypercomplex Quality Assessment of Multi/Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.*, 6(4): 662–665.
- Giuseppe, M.; Davide, C.; Luisa, V.; and Giuseppe, S. 2016. Pansharpening by Convolutional Neural Networks. *Remote Sens.*, 8(7): 594.
- He, X.; Cao, K.; Zhang, J.; Yan, K.; Wang, Y.; Li, R.; Xie, C.; Hong, D.; and Zhou, M. 2025. Pan-mamba: Effective pan-sharpening with state space model. *Inf. Fusion*, 115: 102779.
- He, Y.; Zhu, J.; Li, Y.; Huang, Q.; Wang, Z.; and Yang, K. 2024. Rethinking Remote Sensing CLIP: Leveraging Multimodal Large Language Models for High-Quality Vision-Language Dataset. In *Int. Conf. Neural Inf. Process.*, 417–431. Springer.
- Hou, J.; Cao, Z.; Zheng, N.; Li, X.; Chen, X.; Liu, X.; Cong, X.; Hong, D.; and Zhou, M. 2024. Linearly-evolved transformer for pan-sharpening. In *Proc. ACM Int. Conf. Multimed.*, 1486–1494.
- Kwarteng, P.; and Chavez, A. 1989. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens.*, 55(1): 339–348.
- Lin, H.; Dong, Y.; Ding, X.; Liu, T.; and Liu, Y. 2024. Unsupervised Pan-Sharpener via Mutually Guided Detail Restoration. *Proc. AAAI Conf. Artif. Intell.*, 38(4): 3386–3394.
- Liu, L.; Zhang, J.; Zhou, B.; Lyu, P.; and Cai, Z. 2025. SSA-GAN: Singular Spectrum Analysis-Enhanced Generative Adversarial Network for Multispectral Pansharpening. *Mathematics*, 13(5).
- Ma, J.; Yu, W.; Chen, C.; Liang, P.; Guo, X.; and Jiang, J. 2020. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Inf. Fusion*, 62: 110–120.
- Ma, M.; Zhao, M.; Jiang, Y.; Li, X.; and Zhang, W. 2025. SSFMamba: Spatial-Spectral Fusion State Space Model for Pansharpening. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1–5.
- Mangolini, M.; Ranchin, T.; and Wald, L. 1995. Evaluation de la qualité des images multispectrales à haute résolution

- spatiale dérivées de SPOT. *Rev. Française Photogramm. Télédétection*, 137: 24–29.
- Otazu, X.; Gonzalez-Audicana, M.; Fors, O.; and Nunez, J. 2005. Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods. *IEEE Trans. Geosci. Remote Sens.*, 43(10).
- Peng, S.; Deng, L.; Hu, J.; and Zhuo, Y. 2022. Source-Adaptive Discriminative Kernels based Network for Remote Sensing Pansharpening. In Raedt, L. D., ed., *Proc. Thirty-First Int. Joint Conf. Artif. Intell.*, 1283–1289. ijcai.org.
- Peng, S.; Zhu, X.; Deng, H.; Deng, L.-J.; and Lei, Z. 2024. FusionMamba: Efficient Remote Sensing Image Fusion With State Space Model. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–16.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proc. Int. Conf. Mach. Learn.*, volume 139 of *Proc. Mach. Learn. Res.*, 8748–8763. PMLR.
- Shen, H.; Zhang, B.; Jiang, M.; and Li, J. 2024. Unsupervised Pan-Sharpener Network Incorporating Imaging Spectral Prior and Spatial-Spectral Compensation. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–16.
- Shutao; Yang; and Bin. 2011. A New Pan-Sharpener Method Using a Compressed Sensing Technique. *IEEE Trans. Geosci. Remote Sens.*, 49(2): 738–746.
- Vicinanza, M. R.; Restaino, R.; Vivone, G.; Dalla Mura, M.; and Chanussot, J. 2014. A Pansharpening Method Based on the Sparse Representation of Injected Details. *IEEE Geosci. Remote Sens. Lett.*, 12(1): 180–184.
- Vivanco Cepeda, V.; Nayak, G. K.; and Shah, M. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Adv. Neural Inf. Process. Syst.*, 36: 8690–8701.
- Vivone, G. 2019. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE Trans. Geosci. Remote Sens.*, 57(9): 6421–6433.
- Vivone, G.; Alparone, L.; Chanussot, J.; Mura, M. D.; Garzelli, A.; Licciardi, G. A.; Restaino, R.; and Wald, L. 2015. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.*, 53(5): 2565–2586.
- Vivone, G.; Deng, L.-J.; Deng, S.; Hong, D.; Jiang, M.; Li, C.; Li, W.; Shen, H.; Wu, X.; Xiao, J.-L.; Yao, J.; Zhang, M.; Chanussot, J.; García, S.; and Plaza, A. 2024. Deep Learning in Remote Sensing Image Fusion: Methods, protocols, data, and future perspectives. *IEEE Geosci. Remote Sens. Mag.*, 2–43.
- Vivone, G.; Restaino, R.; and Chanussot, J. 2018. A Regression-Based High-Pass Modulation Pansharpening Approach. *IEEE Trans. Geosci. Remote Sens.*, 56(2): 984–996.
- Wald, L. 2002. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES.
- Wang, J.; Lu, T.; Huang, X.; Zhang, R.; and Feng, X. 2024. Pan-sharpening via conditional invertible neural network. *Inf. Fusion*, 101: 101980.
- Wang, Y.; Lin, Y.; He, X.; Zheng, H.; Yan, K.; Fan, L.; Huang, Y.; and Ding, X. 2025a. Learning Diffusion High-Quality Priors for Pan-Sharpener: A Two-Stage Approach With Time-Aware Adapter Fine-Tuning. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–14.
- Wang, Y.; Lin, Y.; He, X.; Zheng, H.; Yan, K.; Fan, L.; Huang, Y.; and Ding, X. 2025b. Learning Diffusion High-Quality Priors for Pan-Sharpener: A Two-Stage Approach With Time-Aware Adapter Fine-Tuning. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–14.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.
- Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A Deep Network Architecture for Pan-Sharpener. In *Proc. IEEE Int. Conf. Comput. Vis.*, 5449–5457.
- Yuhas, R. H.; Goetz, A. F.; and Boardman, J. W. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *JPL, Summ. Third Annu. JPL Airborne Geosci. Workshop*, volume 1, 1–3.
- Zanella, M.; and Ben Ayed, I. 2024. Low-rank few-shot adaptation of vision-language models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 1593–1603.
- Zeng, H.; Yang, X.; Shen, K.; Li, Y.; Jiang, J.; and Li, F. 2025a. Cross-Modal Contrastive Pansharpening via Uncertainty Guidance. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–14.
- Zeng, H.; Yang, X.; Shen, K.; Li, Y.; Jiang, J.; and Li, F. 2025b. Cross-Modal Contrastive Pansharpening via Uncertainty Guidance. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–14.
- Zhang, J.; Fang, F.; Wang, T.; Zhang, G.; and Song, H. 2025. FrDiff: Framelet-based Conditional Diffusion Model for Multispectral and Panchromatic Image Fusion. *IEEE Trans. Multimed.*, 1–14.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 16816–16825.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9): 2337–2348.
- Zhou, M.; Zheng, N.; He, X.; Hong, D.; and Chanussot, J. 2025. Probing Synergistic High-Order Interaction for Multi-Modal Image Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2): 840–857.
- Zhu, C.; Deng, S.; Zhou, Y.; Deng, L.-J.; and Wu, Q. 2023. QIS-GAN: A lightweight adversarial network with quadtree implicit sampling for multispectral and hyperspectral image fusion. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–15.