Video Can Teach PAN-Sharpening: PSF-Aware Cross-Domain Supervision

Hyun-Ho Kim[1], Munchurl Kim[2], and Jaehyup Lee[3†]

[1] KARI, Daejeon, Republic of Korea
kimhh@kari.re.kr
[2] KAIST, Daejeon, Republic of Korea
mkimee@kaist.ac.kr
[3] Kyungpook National University, Daegu, Republic of Korea
jaehyuplee@knu.ac.kr

Abstract. Most PAN-sharpening networks are trained on synthetically degraded panchromatic (PAN) and multispectral (MS) pairs because real high-resolution MS (HRMS) labels are rarely available. However, real satellite acquisitions often suffer from cross-modal misregistration and sensor-dependent optical responses that are absent from synthetic training pipelines, leading to spectral distortions and double-edge artifacts at test time. This suggests that the key bottleneck lies not only in network design but also in the supervision paradigm. In this work, we show that natural videos can serve as a *spatial teacher* for PAN-sharpening: nearby HR frame pairs provide abundant high-frequency structures and exhibit frame-to-frame shifts that mimic realistic misregistration, enabling strong spatial supervision without HRMS labels. Based on this insight, we propose **ViPS**, a cross-domain training framework under the paradigm of "**Vi**deo can teach **P**AN-**S**harpening". ViPS disentangles supervision sources by learning spatial detail restoration from video-derived pseudo PAN–MS pairs, while enforcing spectral fidelity from real satellite PAN–MS pairs. To align optical degradations across domains, we utilize point spread function (PSF) banks and randomly sample a kernel during training, enabling physically grounded cross-domain learning. Extensive experiments across multiple sensors show that ViPS outperforms very recent state-of-the-art methods under both reduced- and full-resolution settings, while maintaining fast inference.

Keywords: PAN-sharpening · cross-domain supervision · PSF

1 Introduction

Remote sensing (RS) imagery is crucial for a wide range of applications, including urban monitoring, environmental analysis, and infrastructure inspection. Many of these applications demand high-resolution multispectral (HRMS) imagery

† Corresponding author.

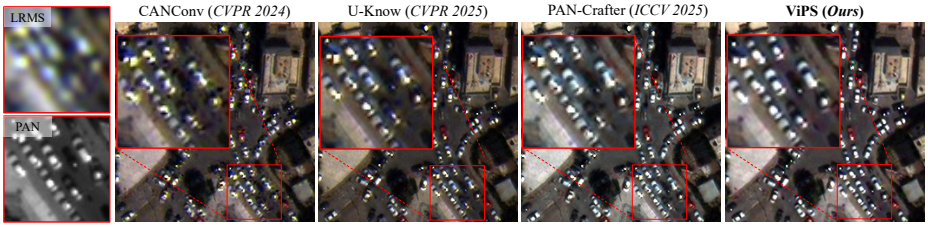


Fig. 1: Visual comparison of PAN-sharpening results on the full-resolution WorldView-3 dataset. Our ViPS result is shown in the rightmost column. Compared with recent baselines, including CANConv [12], U-Know [19], and PAN-Crafter [11], ViPS better preserves fine structures with fewer artifacts. In the highlighted region (red box), ViPS recovers complex layouts more clearly while maintaining spectral spatial details.

that preserves both fine spatial structures and rich spectral information. However, due to inherent constraints of sensor design, a single imaging system cannot simultaneously achieve high spatial resolution and high spectral fidelity. Modern Earth observation satellites therefore provide an HR panchromatic (HRPAN) image and a low-resolution MS (LRMS) observation. PAN-sharpening (PS) aims to fuse these two modalities to reconstruct an HRMS image.

Despite remarkable progress driven by deep learning, a fundamental gap remains between *how* PS models are trained and *how* they are used in practice. Most PS networks are optimized in a fully supervised manner on synthetically degraded PAN–MS pairs constructed by the Wald protocol [31], because real HRMS labels are rarely available. However, satellite PAN–MS acquisitions often exhibit cross-modal misregistration (e.g., inter-sensor offsets, platform motion, and pushbroom acquisition timing) [11, 20], as well as sensor-dependent optical responses that are absent in synthetic training pairs. This train–test mismatch frequently manifests as blurred structures, texture inconsistency, or double-edge artifacts at test time (Fig. 1).

A common remedy is to inject additional spatial priors, such as gradient/edge constraints [8], alignment-aware modules [20], or spatial supervision designs [11]. While effective in certain cases, these approaches can be unreliable: edge-based constraints may fail because PAN edges do not always align with the luminance structures of HRMS, and overly strict spatial alignment can introduce spectral distortion. Moreover, most existing solutions remain *architecture-centric*, in that they mainly compensate for misregistration at the feature level while still being constrained by the realism and diversity of satellite-only supervision. This leaves open a practical question: *where can we obtain scalable and reliable spatial supervision for PS beyond satellite-only training?*

To address this limitation, we propose **ViPS**, a cross-domain learning framework under the paradigm of **"Video can teach PAN-Sharpening"**. **At its core**, ViPS disentangles supervision sources: (i) **video-derived pseudo PAN–MS pairs** provide scalable *spatial supervision* by exposing the model to abundant high-frequency structures and misregistration-relevant correspondence variations, without requiring satellite HRMS labels; (ii) **real satellite PAN–MS pairs** provide reliable *spectral supervision* in the observed LRMS domain.

To align optical degradations across heterogeneous sensors and the video domain, we construct point spread function (PSF) banks [7] and randomly sample kernels during training, yielding a physically grounded degradation operator. Our main contributions are summarized as follows:

- **Video-based training for PS via PSF-bank-aware pseudo supervision.** We introduce a video-derived pseudo PAN–MS construction pipeline that emulates realistic misregistration variations and uses PSF-consistent degradations to teach spatial detail restoration;
- **A multi-sensor training framework with disentangled supervision.** We present a novel training framework applicable across multiple sensors: for each sensor, we use real satellite PAN–MS pairs for spectral supervision and video-derived pseudo pairs for spatial supervision, resulting in improved PAN-sharpening performance;
- **ViPS dataset release.** We curate and will publicly release the ViPS dataset consisting of 10 scenarios, each containing 1000 frame pairs at 3840×2160 resolution, along with scripts for pseudo-pair generation and evaluation. ViPS dataset is publicly available at https://github.com/brachiohyup/ViPS.

2 Related Work

Traditional PAN-sharpening. Traditional PS methods are broadly categorized into component substitution (CS) [6, 9, 27, 30], multi-resolution analysis (MRA) [1, 2, 24, 29], and variational optimization (VO) [3, 5, 13, 33]. CS substitutes an intensity-like component with PAN, MRA injects PAN-driven details across scales, and VO formulates explicit spectral–spatial regularization. Many MRA variants further incorporate sensor optics via modulation transfer function (MTF) and PSF modeling [2,7], which affect degradation behavior in both training and evaluation. Despite their interpretability and efficiency, these methods rely on handcrafted priors and largely linear injection rules, limiting robustness under heterogeneous optics and real acquisition artifacts.

Deep learning-based PAN-sharpening. Recent advancements in PS have been driven by deep learning (DL)-based methods [4, 12, 14, 15, 18, 25, 38, 40]. Most approaches adopt end-to-end regression with CNN backbones [21,23,35–37] and, more recently, Transformer-style architectures to better capture non-local dependencies. Early CNN-based frameworks such as PNN [21] and PanNet [35] establish strong baselines by learning residual and detail injection, while multi-stage refinement and dual back-projection designs, such as S2DBPN [37], further improve high-frequency restoration. To enhance adaptability under diverse scenes, LAGConv [16] introduces content- and context-adaptive operators, and CANConv [12] exploits local-context adaptivity and non-local self-similarity to improve spatial-spectral consistency.

Beyond deterministic regression, diffusion-based models [17, 19, 22, 34, 39] offer a promising alternative by iteratively refining predictions, often producing sharper and more perceptually realistic outputs. However, diffusion-based PS is typically computationally expensive and still commonly relies on synthetic

HRMS supervision, which can limit practicality and real-world robustness. U-Know-DiffPAN [19] introduces an uncertainty-aware teacher–student diffusion framework to emphasize uncertain regions while enhancing details, but the overall paradigm largely remains within satellite-only supervision assumptions.

Misalignment-aware fusion. Real PAN–MS pairs often suffer from misalignment due to inter-sensor offsets, parallax, and platform motion, leading to spectral distortion and double-edge artifacts. To mitigate this, SIPSA [20] incorporates alignment-aware modules, while LAGConv [16] and CANConv [12] use adaptive kernels for implicit matching. PAN-Crafter [11] further enforces modality-consistent alignment via joint reconstruction and alignment-aware attention. Despite these advances, most methods mainly rely on architectural remedies and remain limited by the realism of synthetic training data.

Our position. As illustrated in Fig. 2, ViPS shifts the focus from architectural design to the question of how to provide effective supervision for PAN-sharpening. We disentangle supervision sources: natural videos provide spatial supervision, while real RS PAN–MS pairs provide reliable spectral supervision. The PSF banks align degradations across domains, enabling physically consistent training and mitigating the reliance on RS-only supervision.

3 Method

3.1 Problem Setup and Notation

Given a real satellite observation consisting of an HRPAN–LRMS image pair $\left(\mathbf{I}^h_{pan}, \mathbf{I}^l_{ms}\right)$, PS aims to recover an HRMS image $(\mathbf{I}^h_{ms})$ that inherits the spatial structures from $\mathbf{I}^h_{pan}$ while preserving the spectral fidelity of $\mathbf{I}^l_{ms}$. We set the spatial scale ratio (r) between HR (h) and LR (l) to 4 in all experiments, and denote video-derived pseudo HRPAN–LRMS image pairs as $\left(\bar{\mathbf{I}}^{h,p}_{pan}, \mathbf{I}^{l,p}_{ms}\right)$. Each RS sample is associated with a satellite sensor label (s), and C_s denotes the number of MS bands for s.

3.2 Overview

Fig. 2 illustrates the overall ViPS framework. ViPS *disentangles supervision sources* for PS: (i) **Video branch** provides *spatial supervision* from natural videos, exposing the model to abundant high-frequency structures and pushbroom-like correspondence variations without requiring satellite HRMS labels; (ii) **RS branch** provides reliable *spectral supervision* from real PAN–MS observations.

A key obstacle for cross-domain learning is the optical domain gap between natural cameras and satellite sensors, dominated by sensor-dependent PSFs. ViPS adopts the PSF banks [7] ($\{\mathcal{B}_s\}_{s\in\mathcal{S}}, \mathcal{S} = \{$WorldView-3 (WV3), QuickBird (QB), GaoFen-2 (GF2)$\}$). For each RS sample associated with s, we randomly sample a PSF kernel from the corresponding PSF bank. In addition, when generating video-derived pseudo observations for s, we apply the same $\mathcal{K}_s$ so that the video and RS supervision streams share consistent optical degradation for that sensor.

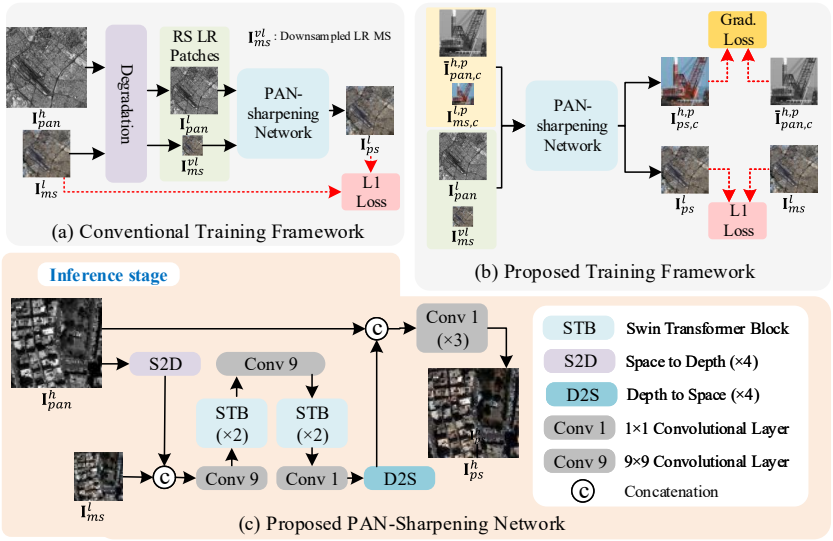


Fig. 2: Overview of the proposed ViPS framework. (a) Prior PS methods typically train only on remote sensing PAN–MS pairs. (b) ViPS disentangles supervision sources: natural videos provide high-quality spatial supervision, while real satellite observations provide spectral supervision via a PSF-consistent objective. (c) Network architecture of ViPS: we instantiate the PS backbone as a hybrid convolution–Transformer network.

3.3 PSF-aware Degradation

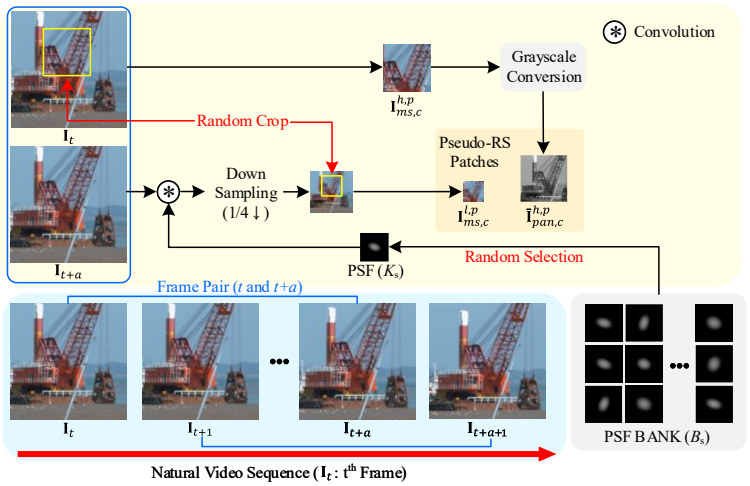
A key challenge in our cross-domain setting is the optical gap between natural video frames and real RS images. In particular, satellite imagery is governed by sensor-dependent PSFs [28], whereas a fixed isotropic blur is often too simplistic to reflect real acquisition characteristics [7, 26]. Motivated by this observation, we introduce a *shared* PSF-aware degradation operator that is used consistently in both the video and RS branches.

PSF bank and random kernel sampling. Real satellite optical responses can be anisotropic and sensor-dependent, and introducing randomized, physically plausible spatial degradations can improve robustness to such variability [7]. Therefore, we define the PSF bank

$$\mathcal{B}_s = \{\mathcal{K}_s^{(m)}\}_{m=1}^{M_s}, \tag{1}$$

where $\mathcal{K}_s^{(m)}$ represents one plausible optical response of s, thereby covering a range of sensor-consistent blur patterns rather than a single fixed degradation model. In all experiments, $\mathcal{K}_s^{(m)} \in \mathbb{R}^{16\times16}$ and we set $M_s = 10{,}000$ for each sensor. All kernels are pre-generated offline following the protocol of [7] and normalized to have unit sum. During training, we uniformly sample

$$\mathcal{K}_s \sim \mathrm{Unif}(\mathcal{B}_s). \tag{2}$$



(a) Proposed Pseudo-RS Patch Generation Process

Fig. 3: Video-derived pseudo HRPAN–LRMS pair construction in ViPS. Given a frame pair $(\mathbf{I}_t, \mathbf{I}_{t+a})$, we form a pseudo HRPAN by grayscale conversion of $\mathbf{I}_t$ and generate a pseudo LRMS by applying channel projection $\mathcal{P}_{C_s}(\cdot)$ followed by the sensor-conditioned PSF-bank degradation $\mathcal{D}(\cdot; \mathcal{K}_s, r)$, where the target sensor s and kernel $\mathcal{K}_s \sim \mathcal{B}_s$ are randomly sampled. We then randomly crop pseudo-RS patches to construct ViPS training samples used as video-derived spatial supervision.

We then instantiate degradation using the sampled kernel. This random sampling increases the diversity of degradation patterns seen during training and prevents the network from overfitting to a single fixed PSF, while preserving physical plausibility. Given an HR multi-channel image ($\mathbf{I}^{HR} \in \mathbb{R}^{H \times W \times C}$) and a PSF kernel ($\mathcal{K}_s$), we define the degradation process ($\mathcal{D}$) to obtain an LR multi-channel image ($\mathbf{I}^{LR} \in \mathbb{R}^{H/r \times W/r \times C}$) as

$$\mathbf{I}^{LR} = \mathcal{D}(\mathbf{I}^{HR}; \mathcal{K}_s, r) = (\mathcal{K}_s * \mathbf{I}^{HR}) \downarrow_r, \tag{3}$$

where $*$ denotes a convolution operator applied independently to each channel, and $\downarrow_r$ denotes downsampling by a factor of r. This formulation follows the standard PS degradation view in which blur and resolution reduction are jointly modeled by a scale-transfer operator.

Cross-domain kernel sharing. The $\mathcal{K}_s$ is used in both branches within each training iteration: (i) to synthesize video-derived pseudo LRMS observations, and (ii) to construct downsampled RS training pairs in the RS branch. By applying the same degradation form across domains, ViPS explicitly aligns the effective blur and frequency attenuation associated with s, making the video and RS supervision physically consistent. This matched degradation is also analogous to the principle used in physically grounded PS data generation, where the same kernel is applied to paired inputs to preserve their cross-domain relationship.

3.4 Video Branch: Spatial Supervision

Motivation. Natural videos provide abundant high-frequency structures at scale. Moreover, satellite sensors typically follow a pushbroom acquisition mechanism, where different scanlines are captured sequentially with small temporal offsets. When the dominant motion is approximately vertical translation and the scene is mostly static, such temporal offsets induce correspondence variations that resemble realistic PAN–MS misregistration. ViPS exploits this property to learn *spatial detail restoration* without satellite HRMS labels.

Pseudo PAN construction. As shown in Fig. 3, we construct a pseudo HRPAN–LRMS pair from a frame pair $(\mathbf{I}_t, \mathbf{I}_{t+a})$. Specifically, we obtain the pseudo HRPAN image $\bar{\mathbf{I}}^{h,p}_{pan}$ by grayscaling $\mathbf{I}_t$:

$$\bar{\mathbf{I}}^{h,p}_{pan} = \mathrm{Gray}(\mathbf{I}_t), \tag{4}$$

where $\mathrm{Gray}(\cdot)$ denotes channel averaging.

Pseudo LRMS generation with PSF bank. To form a pseudo LRMS observation with C_s for s, we first apply a fixed channel projection ($\mathcal{P}_{C_s}(\cdot)$) to the RGB frame $\mathbf{I}_{t+a}$ to produce a C_s-channel input, and then apply the PSF-bank degradation as

$$\mathbf{I}^{l,p}_{ms} = \mathcal{D}\big(\mathcal{P}_{C_s}(\mathbf{I}_{t+a}); \mathcal{K}_s, r\big), \tag{5}$$

where $\mathcal{K}_s$ is sampled from $\mathcal{B}_s$ (Eq. 2). Here, $\mathcal{P}_{C_s}(\cdot)$ is used only to match the required band dimensionality of sensor s for input compatibility; it is not intended to mimic the true spectral response of the satellite sensor.

Spatial learning with pseudo pairs. Given the pseudo pair $(\bar{\mathbf{I}}^{h,p}_{pan}, \mathbf{I}^{l,p}_{ms})$, the PS network (f_θ) predicts pseudo HRMS:

$$\hat{\mathbf{I}}^{h,p}_{ms} = f_\theta(\bar{\mathbf{I}}^{h,p}_{pan}, \mathbf{I}^{l,p}_{ms}). \tag{6}$$

Spatial loss. We supervise structure by matching gradient magnitudes in a luminance space:

$$\bar{\mathbf{I}}^{h,p}_{ms} = \mathrm{Gray}(\hat{\mathbf{I}}^{h,p}_{ms}), \quad \mathcal{L}_{spa} = \Big\| \, |\nabla \bar{\mathbf{I}}^{h,p}_{ms}| - |\nabla \bar{\mathbf{I}}^{h,p}_{pan}| \, \Big\|_1, \tag{7}$$

where ∇ is a gradient operator. This formulation is empirically robust to small cross-modal misregistration, as it avoids strict pixel-wise edge correspondence while still enforcing sharp structural consistency.

3.5 RS Branch: Spectral Supervision

Motivation. Real satellite PAN–MS pairs provide reliable spectral information but lack HRMS ground truth. To exploit the observed LRMS as a spectral target, we construct downsampled RS training pairs using the same sensor-conditioned degradation employed in the video branch. Given a real RS pair $(\mathbf{I}^h_{pan}, \mathbf{I}^l_{ms})$ and $\mathcal{K}_s$, we form

$$\mathbf{I}^l_{pan} = \mathcal{D}(\mathbf{I}^h_{pan}; \mathcal{K}_s, r), \qquad \mathbf{I}^{vl}_{ms} = \mathcal{D}(\mathbf{I}^l_{ms}; \mathcal{K}_s, r). \tag{8}$$

At this downsampled scale, LR PAN ($\mathbf{I}^l_{pan}$) and very low-resolution MS ($\mathbf{I}^{vl}_{ms}$) play the roles of PAN and LRMS inputs, respectively. Subsequently, we define the spectral loss as

$$\mathcal{L}_{spe} = \left\| f_\theta(\mathbf{I}^l_{pan}, \mathbf{I}^{vl}_{ms}) - \mathbf{I}^l_{ms} \right\|_1. \tag{9}$$

This loss encourages the reconstructed MS image at the reduced-resolution scale to remain spectrally consistent with the observed LRMS, which serves as the supervisory target.

3.6 PAN-Sharpening Network

As illustrated in Fig. 2-(c), in our proposed ViPS framework, we introduce a lightweight hybrid convolution–Transformer backbone to balance efficiency and representational capacity. Given $\mathbf{I}^h_{pan}$ and $\mathbf{I}^l_{ms}$, we first apply a space-to-depth (S2D) transform to $\mathbf{I}^h_{pan}$ so that its spatial resolution matches that of $\mathbf{I}^l_{ms}$. The S2D-transformed PAN feature and $\mathbf{I}^l_{ms}$ are concatenated channel-wise and processed by a 9×9 convolution followed by stacked Swin Transformer blocks to capture both local structures and non-local spatial–spectral interactions. A second 9×9 convolution and another set of Swin Transformer blocks further refine the fused representation. Finally, we project features with a 1×1 convolution and upsample to PAN resolution via a depth-to-space (D2S) operation. The upsampled features are concatenated with the original PAN image to explicitly inject fine-grained textures, and three successive 1×1 convolutions produce the final HRMS output $\hat{\mathbf{I}}^h_{ms}$.

3.7 Overall Objective and Training Strategy

We optimize the PS network with:

$$\mathcal{L} = \lambda_{spa}\mathcal{L}_{spa} + \lambda_{spe}\mathcal{L}_{spe}. \tag{10}$$

We alternate RS mini-batches and video mini-batches (1:1 by default). For each RS mini-batch from s, we randomly sample $\mathcal{K}_s$ from $\mathcal{B}_s$ and use it to construct the downsampled RS training pair used by the spectral loss. We then construct the subsequent video mini-batch with the same target s (thus the same band dimension C_s) and apply the same $\mathcal{K}_s$ to generate pseudo LRMS. This sensor-matched scheduling ensures that video and RS supervision share consistent optics within each sensor, while different sensors use different kernels sampled from their own banks.

4 Proposed ViPS Dataset

4.1 Motivation and Design Principles

As mentioned earlier, our ViPS is designed under the paradigm that videos can teach PAN-sharpening. Specifically, natural videos offer scalable spatial supervision, while real remote sensing PAN–MS data can provide reliable spectral

supervision. Importantly, our goal is *not* to perfectly match satellite radiometry with videos. Instead, our ViPS is tailored to act as a high-quality spatial teacher by providing: (i) abundant high-frequency structures, and (ii) *frame-to-frame shifts that mimic realistic misregistration*, both of which are directly relevant to cross-modality matching and misregistration robustness.

4.2 Data Curation Protocol

The ViPS dataset consists of high-quality video sequences captured under diverse scenes and illumination conditions using a high-speed camera at 4K resolution and 1,000 frames per second (FPS). To emulate a misregistration-relevant setting, we record sequences in which the camera motion is dominated by translation with minimal rotation, while the scene remains mostly static. We discard clips with severe motion blur, abrupt exposure changes, strong rolling-shutter artifacts, or dynamic foreground motions that break frame-to-frame correspondence. We follow these guidelines to curate the ViPS dataset: (i) translation motion with minimal rotation; (ii) a high frame rate (1,000 FPS) to keep the inter-frame displacement small, analogous to real satellite image acquisition; (iii) diverse scene types (e.g., urban areas, vegetation, and roads) to include the variety of land-cover categories commonly observed in RS.

Overall, ViPS contains **10** scenarios, each captured as a video clip of approximately 5 seconds at **3,840×2,160** resolution and **1,000 FPS**, resulting in about **5,000** frames per scenario. We adopt a 4K, 1,000-FPS camera for two reasons: the 4K resolution preserves fine spatial structures and enables diverse, high-quality crops, while the high frame rate approximates the high line-rate regime of pushbroom satellite imaging, reduces motion blur, and keeps frame-to-frame displacement small and controllable. Although this setup does not explicitly reproduce true scanline-wise pushbroom sensing, it provides a practical approximation of the low-blur, temporally dense acquisition setting needed to construct frame pairs with pushbroom-relevant correspondence variations. From each scenario, we construct frame pairs $(\mathbf{I}_t, \mathbf{I}_{t+a})$ by selecting a temporal offset a, which controls the magnitude of frame-to-frame correspondence variation. In our experiments, we examine $a \in \{1, 5, 10, 20\}$ and construct **1,000** frame pairs per scenario. As shown in the ablation study (Sec. 5.5), $a = 10$ provides the best PAN-sharpening performance. Accordingly, we use $a = 10$ in all main experiments and plan to publicly release the ViPS dataset curated under the $a = 10$ setting, containing **1,000** pairs per scenario (**10,000** pairs in total). Example frames and further details of the ViPS dataset construction pipeline are provided in the *suppl*.

4.3 Pre-processing and Patch Construction

We extract frames and apply quality control to remove frames with severe blur or unstable exposure. For training, we use cropped patches of size 64×64 at the pseudo-PAN resolution. To control the amount of correspondence variation, we sample frame pairs with a short temporal offset a ($a = 10$; larger a yields

stronger translation shifts). We set $a = 10$ in all main experiments based on the ablation study over $a \in \{1, 5, 10, 20\}$ (Sec. 5.5).

4.4 Pseudo PAN–MS Pair Generation

Given a frame pair $(\mathbf{I}_t, \mathbf{I}_{t+a})$ sampled from the same scenario, we form a pseudo HRPAN image by grayscale conversion:

$$\bar{\mathbf{I}}^{h,p}_{pan} = \mathrm{Gray}(\mathbf{I}_t), \tag{11}$$

and generate a pseudo LRMS observation by PSF-bank-aware degradation:

$$\mathbf{I}^{l,p}_{ms} = \mathcal{D}\big(\mathcal{P}_{C_s}(\mathbf{I}_{t+a}); \mathcal{K}_s, r\big), \qquad r = 4. \tag{12}$$

Notably, using $\mathcal{K}_s$ is crucial because it aligns the degradation models across domains, leading to physically grounded cross-domain learning.

Upon acceptance, we plan to publicly release ViPS with a fixed temporal offset $a = 10$, curating 1,000 pseudo PAN–MS pairs per scenario (10,000 pairs total). The release will include the selected pairs and the detailed metadata, while further curation details and sample pairs are described in the *suppl.*

5 Experiments

5.1 Datasets

Video-derived pseudo PAN–MS pairs (ViPS dataset). We construct video-derived pseudo supervision using the proposed ViPS dataset (Sec. 4). Each sample is formed from a frame pair $(\mathbf{I}_t, \mathbf{I}_{t+a})$: we obtain pseudo PAN as $\mathrm{Gray}(\mathbf{I}_t)$ and generate pseudo LRMS by applying $\mathcal{P}_{C_s}(\cdot)$ followed by $\mathcal{D}(\cdot; \mathcal{K}_s, r)$ with $r = 4$. We set C_s to match the band dimension of the target satellite sensor (e.g., C_s=8 for WV3 and C_s=4 for GF2 and QB) for architectural compatibility. This projection is *not* intended to approximate satellite radiometry; it only facilitates video-derived *spatial* supervision.

Remote sensing datasets and evaluation protocols. For real satellite imagery, we use the PAN-Collection benchmark [10], which provides multi-sensor PAN–MS pairs, including WV3, GF2, and QB, with a scale ratio of $r = 4$. Unless otherwise specified, we follow the official train/test splits and evaluation code. Each dataset provides reduced-resolution (RR) and full-resolution (FR) test protocols. The RR test images have a PAN spatial size of 256×256, while the FR test images have a larger PAN spatial size of 512×512.

5.2 Implementation Details

All models are implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We use a downsampling factor $r = 4$ in all experiments.

Table 1: Quantitative comparison on WorldView-3 under full-resolution and reduced-resolution protocols. We report HQNR with its distortion terms (D_s, D_λ) for FR, and standard full-reference metrics for RR. Best and second-best results are marked in **red** and blue, respectively. Inference time and FLOPs are measured on a 256×256 HRMS target (batch size 1) under the RR setting.

Methods	Pubs.	Full-Resolution			Reduced-Resolution						Infer.	FLOPs
		HQNR↑	D_s ↓	D_λ ↓	ERGAS↓	SCC↑	SAM↓	Q↑	PSNR↑	SSIM↑	Time(s)↓	(G)↓
PanNet [35]	ICCV 2017	0.918	0.049	0.035	2.538	0.979	3.402	0.913	36.148	0.966	0.001	5.17
MSDCNN [36]	JSTARS 2018	0.924	0.050	0.028	2.489	0.979	3.300	0.914	36.329	0.967	0.002	14.96
FusionNet [32]	ICCV 2021	0.920	0.053	0.029	2.428	0.981	3.188	0.916	36.569	0.968	0.002	5.13
LAGConv [16]	AAAI 2022	0.915	0.055	0.033	2.380	0.981	3.153	0.916	36.732	0.970	0.004	8.43
S2DBPN [37]	TGRS 2023	0.946	0.030	0.025	2.245	0.985	3.019	0.917	37.216	0.972	0.005	158.94
PanDiff [22]	TGRS 2023	0.952	0.034	**0.014**	2.276	0.984	3.058	0.913	37.029	0.971	2.955	62.07
DCPNet [38]	TGRS 2024	0.923	0.036	0.043	2.301	0.984	3.083	0.915	37.009	0.972	0.109	105.40
TMDiff [34]	TGRS 2024	0.924	0.059	0.018	2.151	0.986	2.885	0.915	37.477	0.973	9.997	1284.42
CANConv [12]	CVPR 2024	0.951	0.030	0.020	2.163	0.985	2.927	0.918	37.441	0.973	0.451	52.21
U-Know [19]	CVPR 2025	0.955	0.029	0.016	2.046	0.988	2.797	0.922	37.934	0.976	-	-
PAN-Crafter [11]	ICCV 2025	0.958	0.027	0.016	2.040	0.988	2.787	0.922	37.956	0.976	0.009	79.03
ViPS (Ours)	-	**0.959**	**0.026**	0.016	**2.021**	**0.989**	**2.770**	**0.923**	**38.021**	**0.978**	0.003	19.50

Table 2: Quantitative comparison on GaoFen-2 and QuickBird under full-resolution and reduced-resolution protocols. FR reports HQNR and D_s, and RR reports standard full-reference metrics. Best and second-best values among the quality metrics are marked in **red** and blue, respectively.

Methods	**GaoFen-2**						**QuickBird**					
	Full-Resolution		Reduced-Resolution				Full-Resolution		Reduced-Resolution			
	HQNR↑	D_s ↓	ERGAS↓	SCC↑	SAM↓	PSNR↑	HQNR↑	D_s ↓	ERGAS↓	SCC↑	SAM↓	PSNR↑
PanNet [35]	0.929	0.052	1.038	0.975	1.050	39.197	0.851	0.092	4.856	0.966	5.273	35.563
MSDCNN [36]	0.898	0.079	0.862	0.983	0.946	40.730	0.888	0.058	4.074	0.977	4.828	37.040
FusionNet [32]	0.865	0.105	0.960	0.980	0.971	39.866	0.853	0.079	4.183	0.975	4.892	36.821
LAGConv [16]	0.895	0.078	0.816	0.985	0.886	41.147	0.892	**0.035**	3.845	0.980	4.682	37.565
S2DBPN [37]	0.935	0.046	0.686	0.990	0.772	42.686	0.908	0.036	3.956	0.980	4.849	37.314
PanDiff [22]	0.936	0.045	0.674	0.990	0.767	42.827	0.919	0.055	3.723	0.982	4.611	37.842
DCPNet [38]	0.953	0.024	0.724	0.988	0.806	42.312	0.880	0.073	3.618	0.983	4.420	38.079
TMDiff [34]	0.942	0.030	0.754	0.988	0.764	41.896	0.901	0.068	3.804	0.981	4.627	37.642
CANConv [12]	0.919	0.063	0.653	0.991	0.722	43.166	0.893	0.070	3.740	0.982	4.554	37.795
U-Know [19]	0.944	0.037	0.548	0.993	0.624	44.585	0.931	**0.035**	3.500	0.984	4.337	38.361
PAN-Crafter [11]	0.964	0.017	0.522	**0.994**	0.596	45.076	0.920	0.039	3.570	0.984	4.426	38.195
ViPS (Ours)	**0.976**	**0.016**	**0.519**	**0.994**	**0.581**	**45.122**	**0.932**	0.036	**3.464**	**0.985**	**4.308**	**38.451**

Training patches and augmentation. For the RS branch, we utilize patches of size $64 \times 64 \times 1$ from $\mathbf{I}^{l}_{pan}$ and the corresponding $16 \times 16 \times C_{\text{ms}}$ patches from $\mathbf{I}^{vl}_{ms}$, where $C_{\text{ms}} = 4$ for QB/GF2 and $C_{\text{ms}} = 8$ for WV3. For the video branch, we randomly crop 64×64 patches from frames and construct pseudo PAN–MS pairs following Sec. 3.4 and Sec. 4. We apply standard augmentations, including random horizontal/vertical flips and 90° rotations.

PSF bank and sampling. We use PSF banks with $M_s = 10{,}000$ kernels per sensor, each with 16×16 support (Sec. 3.3). For each RS mini-batch from sensor s, we randomly sample $\mathcal{K}_s$ from the corresponding bank and use it to construct the downsampled RS training pair used by the spectral loss; for the subsequent video mini-batch, we use the same sensor s and apply the same $\mathcal{K}_s$ to generate pseudo LRMS, aligning degradations across domains.

Optimization and efficiency. We optimize the PS model using AdamW with learning rate 1×10^{-4} and weight decay 0.01, and train for 10^5 iterations per dataset while alternating RS and video mini-batches with a 1:1 ratio. We set the loss weights to λ_{spa}=10 and λ_{spe}=1 for all experiments.

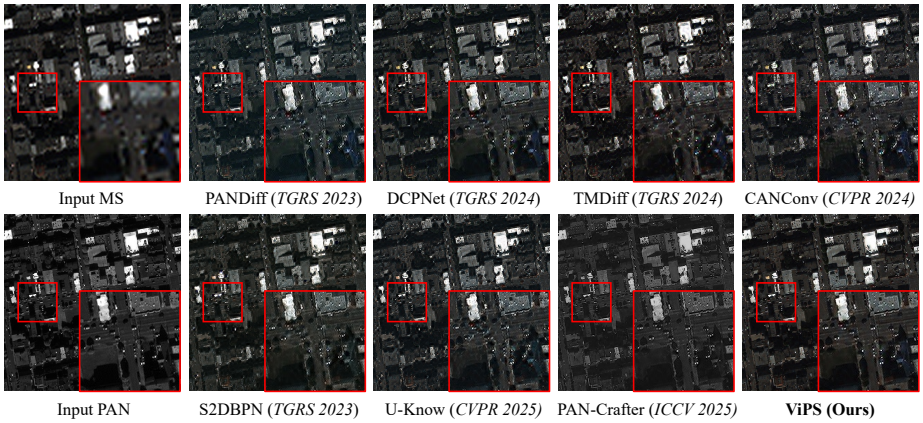


Fig. 4: Visual comparison of PS results on the FR QB dataset. The leftmost columns show the input LRMS and HRPAN images, and the remaining columns present the outputs of very recent SOTA methods and our proposed ViPS. Compared to prior approaches, ViPS better preserves the fine details and spectral tones under real FR settings.

5.3 Quantitative Comparison

Tab. 1 and Tab. 2 report quantitative comparisons under both FR and RR protocols. All metrics are computed with the official PAN-Collection evaluation code, and we use official implementations whenever available.

Full-resolution evaluation. Under the FR setting, ViPS achieves the best HQNR on all three sensors. On WV3, ViPS attains the top HQNR (0.959) with the lowest spatial distortion D_s (0.026) while maintaining strong spectral consistency (D_λ=0.016). Similar trends hold on GF2 and QB, where ViPS again achieves the best HQNR with competitive FR quality.

Reduced-resolution evaluation. Under the RR protocol, ViPS consistently ranks first (or tied first) across standard full-reference metrics. On WV3, ViPS achieves the best ERGAS/SAM/Q/PSNR/SSIM, indicating improvements in both spectral fidelity and overall reconstruction quality. On GF2 and QB, ViPS again delivers the top ERGAS/SAM and achieves the best PSNR, demonstrating robust gains across sensors.

Efficiency. As reported in Tab. 1, ViPS remains lightweight (0.003s, 19.5G FLOPs on a 256×256 target with batch size 1) while outperforming state-of-the-art (SOTA) methods.

5.4 Qualitative Results

Fig. 4 presents representative FR qualitative results on the QB dataset. The leftmost columns show the input LRMS and HRPAN images, followed by the outputs of very recent baselines, including CANConv [12], U-Know [19], and PAN-Crafter [11]. Overall, ViPS produces sharper and more coherent spatial

Table 3: Ablation on WorldView-3 analyzing key ViPS design components and its transferability to a different PS backbone (PanNet). "PSF-bank degrad." indicates the use of the same sensor-conditioned PSF-bank operator $\mathcal{D}(\cdot; \mathcal{K}_s, r)$ in both RS and video branches. "Ran. Sel." denotes random sampling of PSF kernels during training. †PanNet baseline results are taken from Tab. 1.

Variant	Components			Full-Resolution (WV3)			Reduced-Resolution (WV3)		
	$\mathcal{L}_{spa}$ (Video)	PSF-bank degrad.	Ran. Sel.	HQNR↑	$D_s\downarrow$	$D_\lambda\downarrow$	ERGAS↓	SAM↓	PSNR↑
(A) RS-only		✓	✓	0.949	0.036	**0.015**	2.420	2.992	37.118
(B) w/o PSF-bank	✓		–	0.956	0.028	0.016	2.235	2.892	37.277
(C) w/o Ran. Sel.	✓	✓		0.948	0.029	0.023	2.213	3.190	37.613
(D) ViPS (full)	✓	✓	✓	**0.959**	**0.026**	0.016	**2.020**	**2.770**	**38.021**
PanNet (baseline)†	–	–	–	0.918	0.049	0.035	2.538	3.402	36.148
PanNet + ViPS Super.	✓	✓	✓	0.954	0.028	0.018	2.217	2.814	37.229

structures such as building boundaries, thin road markings, and small vehicles. Moreover, ViPS exhibits smaller spectral distortions than prior methods, thanks to spectral supervision from remote sensing PAN–MS pairs, improving spatial-spectral consistency under real FR settings. These visual improvements align with the quantitative gains in D_s and HQNR; additional comparisons and zoom-in versions are provided in the *suppl.*

5.5 Ablation Studies

We ablate key design choices of ViPS on WV3 under both FR and RR settings. In particular, we examine (i) whether adding video-derived pseudo supervision improves a satellite-only baseline, (ii) whether PSF-bank-aware degradation is necessary for stable cross-domain mixing, and (iii) how sensitive the ViPS dataset is to the temporal offset a used to sample video frame pairs $(\mathbf{I}_t, \mathbf{I}_{t+a})$.

Effect of video-derived pseudo supervision. Tab. 3 shows that video-derived pseudo supervision consistently improves an RS-only baseline. Compared to (A) RS-only, (D) ViPS (full) achieves higher HQNR and markedly lowers D_s under the FR protocol, indicating reduced spatial artifacts while keeping D_λ comparable. Under the RR protocol, ViPS also improves ERGAS/SAM/PSNR simultaneously, suggesting that the video branch strengthens spatial detail recovery without compromising spectral fidelity.

Role of PSF-bank-aware degradation and random PSF sampling. We further ablate optics alignment and kernel diversity in Tab. 3. (B) w/o PSF-bank removes the shared sensor-conditioned PSF-bank degradation between the RS and video branches, which degrades both FR and RR performance, implying that cross-domain supervision benefits from matched optics. (C) w/o Ran. Sel. uses a fixed PSF kernel (no random sampling), which increases spectral distortion (higher D_λ) and worsens RR spectral quality (higher SAM). Overall, sharing a sensor-conditioned PSF-bank degradation and randomly sampling kernels are both important for physically consistent and stable cross-domain supervision.

Generalizability to a different PS backbone. ViPS is not tied to a specific PS architecture. As shown in Tab. 3, training PanNet with **ViPS Super.** (our video–RS supervision framework with sensor-conditioned degradations) consistently improves over the PanNet baseline under both FR and RR settings, indicating that ViPS generalizes to different PS backbones.

Table 4: Sensitivity analysis on the temporal offset a used to construct ViPS frame pairs $(\mathbf{I}_t, \mathbf{I}_{t+a})$, evaluated on WorldView-3. All other ViPS components are kept fixed.

Variant	a	Full-Resolution (WV3)			Reduced-Resolution (WV3)		
		HQNR↑	$D_s\downarrow$	$D_\lambda\downarrow$	ERGAS↓	SAM↓	PSNR↑
(A) + ViPS (full)	1	0.954	0.028	**0.016**	2.248	2.814	37.359
(B) + ViPS (full)	5	0.957	0.027	**0.016**	2.239	2.782	37.808
(C) + ViPS (full)	**10**	**0.959**	**0.026**	**0.016**	**2.020**	**2.770**	**38.021**
(D) + ViPS (full)	20	0.958	0.027	**0.016**	2.539	2.994	37.013

Sensitivity to temporal offset a (ViPS dataset setting). Tab. 4 evaluates the proposed ViPS dataset construction by varying the frame interval a used to sample video pairs $(\mathbf{I}_t, \mathbf{I}_{t+a})$, while keeping all other ViPS components fixed. Smaller offsets (a=1, 5) provide limited inter-frame shifts and yield moderate gains, whereas an overly large offset (a=20) reduces correspondence reliability and degrades RR metrics (e.g., ERGAS and PSNR). Overall, a=10 provides the best trade-off in our setting and is used as the default throughout the paper.

6 Conclusion

In this paper, we introduce **ViPS**, a novel supervision-centric and *optics-aware* training paradigm for PAN-sharpening, built on the premise that *"Video can teach PAN-sharpening."* At its core, ViPS *disentangles supervision sources*: (i) video-derived pseudo PAN–MS pairs provide scalable *spatial* supervision by offering abundant high-frequency structures and realistic frame-to-frame shifts that mimic misregistration, without requiring satellite HRMS ground truth; (ii) real RS PAN–MS pairs provide reliable *spectral* cues, which we exploit via an LRMS-domain self-consistency objective.

To align degradations between natural videos and RS observations across heterogeneous sensors, ViPS adopts a unified multi-sensor PSF bank. Specifically, we randomly sample a kernel during training and apply it consistently to both video pseudo-LRMS synthesis and RR RS-pair construction, resulting in matched and physically plausible degradations. Extensive experiments on multiple sensors under both RR and FR protocols show that ViPS significantly and consistently outperforms very recent SOTA methods, including U-Know [19] and PAN-Crafter [11], in terms of spatial detail preservation and spectral fidelity, while maintaining fast inference.

Finally, we curated the **ViPS dataset** and will publicly release the data, splits, and generation scripts, enabling scalable video-derived spatial supervision in a reproducible manner. Looking forward, promising directions include extending the spatial teacher beyond *pushbroom-like* acquisition, incorporating richer PSF priors, and generalizing the proposed cross-domain supervision principle to other multi-modal fusion problems. More broadly, our results indicate that *readily available* natural videos can serve as a practical and scalable supervision source for remote sensing, providing a first step toward broader cross-domain, optics-aware learning for robust image fusion.

Acknowledgments

This work was partially supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government [Ministry of Science and ICT (Information and Communications Technology)] (Project Number: RS-2024-00338513, Project Title: AI-based Computer Vision Study for Satellite Image Processing and Analysis). And, this work was also supported by the Nurturing Global Technical Experts for Performance Evaluation of Multimodal Content Copyright Core Technologies Project through the Korea Creative Content Agency (KOCCA), funded by the Ministry of Culture, Sports and Tourism (MCST) and the Korea Creative Content Agency for Culture Technology (KCTIEP) (Project No. RS-2026-2552393).

References

1. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A.: Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. IEEE Transactions on geoscience and remote sensing **40**(10), 2300–2312 (2002)
2. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M.: Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. Photogrammetric Engineering & Remote Sensing **72**(5), 591–596 (2006)
3. Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B.: A variational model for p+ xs image fusion. International Journal of Computer Vision **69**, 43–58 (2006)
4. Bandara, W.G.C., Patel, V.M.: Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1767–1777 (2022)
5. Cao, X., Chen, Y., Cao, W.: Proximal pannet: A model-based deep network for pansharpening. In: Proceedings of the AAAI conference on artificial intelligence. pp. 176–184 (2022)
6. Carper, W., Lillesand, T., Kiefer, R., et al.: The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. Photogrammetric Engineering and remote sensing **56**(4), 459–467 (1990)
7. Chen, L., Vivone, G., Nie, Z., Chanussot, J., Yang, X.: Spatial data augmentation: Improving the generalization of neural networks for pansharpening. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–11 (2023). https://doi.org/10.1109/TGRS.2023.3262262
8. Choi, J.S., Kim, Y., Kim, M.: S3: A spectral-spatial structure loss for pansharpening networks. IEEE Geoscience and Remote Sensing Letters **17**(5), 829–833 (2019)
9. Choi, J., Yu, K., Kim, Y.: A new adaptive component-substitution-based satellite image fusion by using partial replacement. IEEE transactions on geoscience and remote sensing **49**(1), 295–309 (2010)
10. Deng, L.J., Vivone, G., Paoletti, M.E., Scarpa, G., He, J., Zhang, Y., Chanussot, J., Plaza, A.: Machine learning in pansharpening: A benchmark, from shallow to deep networks. IEEE Geoscience and Remote Sensing Magazine **10**(3), 279–315 (2022)
11. Do, J., Kim, S., Youk, G., Lee, J., Kim, M.: Pan-crafter: Learning modality-consistent alignment for pan-sharpening. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4242–4252 (October 2025)

12. Duan, Y., Wu, X., Deng, H., Deng, L.J.: Content-adaptive non-local convolution for remote sensing pansharpening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27738–27747 (2024)
13. Fu, X., Lin, Z., Huang, Y., Ding, X.: A variational pan-sharpening with local gradient constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10265–10274 (2019)
14. Hou, J., Cao, Q., Ran, R., Liu, C., Li, J., Deng, L.j.: Bidomain modeling paradigm for pansharpening. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 347–357 (2023)
15. Hou, J., Cao, Z., Zheng, N., Li, X., Chen, X., Liu, X., Cong, X., Hong, D., Zhou, M.: Linearly-evolved transformer for pan-sharpening. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1486–1494 (2024)
16. Jin, Z.R., Zhang, T.J., Jiang, T.X., Vivone, G., Deng, L.J.: Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In: Proceedings of the AAAI conference on artificial intelligence. pp. 1113–1121 (2022)
17. Kim, H.H., Kim, M.: A dual-decoder-vae-based latent diffusion model for pan-sharpening. IEEE Geoscience and Remote Sensing Letters **22**, 1–5 (2025). https://doi.org/10.1109/LGRS.2025.3615226
18. Kim, H.H., Kim, M.: Fbs-ps: Fully band-separable pan-sharpening considering the physical characteristics of electro-optical sensors. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–20 (2025). https://doi.org/10.1109/TGRS.2024.3523865
19. Kim, S., Do, J., Lee, J., Kim, M.: U-know-diffpan: An uncertainty-aware knowledge distillation diffusion framework with details enhancement for pan-sharpening. arXiv preprint arXiv:2412.06243 (2024)
20. Lee, J., Seo, S., Kim, M.: Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10166–10174 (2021)
21. Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G.: Pansharpening by convolutional neural networks. Remote Sensing **8**(7), 594 (2016)
22. Meng, Q., Shi, W., Li, S., Zhang, L.: Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–17 (2023)
23. O'shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 (2015)
24. Otazu, X., González-Audícana, M., Fors, O., Núñez, J.: Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods. IEEE Transactions on Geoscience and Remote Sensing **43**(10), 2376–2385 (2005)
25. Peng, S., Zhu, X., Deng, H., Deng, L.J., Lei, Z.: Fusionmamba: Efficient remote sensing image fusion with state space model. IEEE Transactions on Geoscience and Remote Sensing (2024)
26. Pittelkau, M.E., McKinley, W.G.: Optical transfer functions, weighting functions, and metrics for images with two-dimensional line-of-sight motion. Optical Engineering **55**(6), 063108–063108 (2016)
27. Rahmani, S., Strait, M., Merkurjev, D., Moeller, M., Wittman, T.: An adaptive ihs pan-sharpening method. IEEE Geoscience and Remote Sensing Letters **7**(4), 746–750 (2010)
28. Schowengerdt, R.A.: Remote sensing: models and methods for image processing. Elsevier (2006)

29. Shah, V.P., Younan, N.H., King, R.L.: An efficient pan-sharpening method via a combined adaptive pca approach and contourlets. IEEE transactions on geoscience and remote sensing **46**(5), 1323–1335 (2008)
30. Shensa, M.J.: The discrete wavelet transform: wedding the a trous and mallat algorithms. IEEE Transactions on signal processing **40**(10), 2464–2482 (1992)
31. Wald, L.: Quality of high resolution synthesised images: Is there a simple criterion? In: Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images". pp. 99–103. SEE/URISCA (2000)
32. Wu, X., Huang, T.Z., Deng, L.J., Zhang, T.J.: Dynamic cross feature fusion for remote sensing pansharpening. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14687–14696 (2021)
33. Wu, Z.C., Huang, T.Z., Deng, L.J., Hu, J.F., Vivone, G.: Vo+ net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2021)
34. Xing, Y., Qu, L., Zhang, S., Feng, J., Zhang, X., Zhang, Y.: Empower generalizability for pansharpening through text-modulated diffusion model. IEEE Transactions on Geoscience and Remote Sensing (2024)
35. Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J.: Pannet: A deep network architecture for pan-sharpening. In: Proceedings of the IEEE international conference on computer vision. pp. 5449–5457 (2017)
36. Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L.: A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **11**(3), 978–989 (2018)
37. Zhang, K., Wang, A., Zhang, F., Wan, W., Sun, J., Bruzzone, L.: Spatial-spectral dual back-projection network for pansharpening. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023)
38. Zhang, Y., Yang, X., Li, H., Xie, M., Yu, Z.: Dcpnet: a dual-task collaborative promotion network for pansharpening. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
39. Zhong, Y., Wu, X., Deng, L.J., Cao, Z., Dou, H.X.: Ssdiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. Advances in Neural Information Processing Systems **37**, 77962–77986 (2024)
40. Zhu, Z., Cao, X., Zhou, M., Huang, J., Meng, D.: Probability-based global cross-modal upsampling for pansharpening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14039–14048 (2023)