

Deep Learning Architectures for Landsat 9 Pan-Sharpening

Mehmet Toğa ✉

Lecturer, Yozgat Bozok University, Yozgat, Türkiye

Received: 25.08.2026

Accepted: 02.09.2026

Published: 09.10.2026

<https://doi.org/10.54414/OSFI5256>

Copyright: © 2026 by the authors.
Licensee: Journal of Smart Technologies and Computational Systems, Western Caspian University, Baku, Azerbaijan. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Abstract: The 30 m sampling of the Landsat 9 Operational Land Imager-2 (OLI-2) multispectral channels (B2, B3, B4) and the 15 m sampling of the panchromatic channel (B8) bound the spatial resolution at which surface reflectance can be quantitatively retrieved, including over topographically complex, mountainous, and agricultural landscapes where field- and slope-scale features are often smaller than the native multispectral pixel. Five convolutional architectures—DCNN, PanNet, ResASPP, DenseNet, and Attention-Sharpen—were benchmarked for reconstructing the multispectral cube at panchromatic resolution, following the reduced-resolution variant of Wald's protocol, with the multispectral stack degraded by 1/4 and the panchromatic channel by 1/2. A 2048×2048 extent of scene LC09_L1TP_175034_20251109 was partitioned into 1024 non-overlapping 64×64 patches. Each network was trained for 20 epochs with the L1 loss, Adam optimisation ($\text{lr} = 1 \times 10^{-3}$), and a batch size of 8. Quality was assessed with four full-reference metrics (PSNR, SSIM, SAM, ERGAS) and three no-reference metrics (QNR, $D\lambda$, D_s). PanNet, by restricting its learning manifold to the high-frequency residual of the upsampled multispectral cube, attained PSNR = 46.47 dB, SSIM = 0.98, SAM = 0.617°, ERGAS = 1.394, and QNR = 0.9921, outperforming all baselines. Pairwise Wilcoxon signed-rank tests rejected the null hypothesis of equal performance against each baseline at $p < 0.05$. A timing analysis exposed an approximately sevenfold gap between the fastest network (DenseNet, 1.49 s/epoch) and the heaviest (Attention-Sharpen, 10.98 s/epoch), motivating an application-conditioned model-selection guideline for operational Earth-observation pipelines serving high-resolution land and landscape monitoring.

Keywords: pan-sharpening; Landsat 9 OLI-2; deep learning; Wald's protocol; residual high-pass learning; spectral fidelity; landscape monitoring

1. Introduction

The OLI-2 push-broom radiometer on board Landsat 9 acquires the blue (B2, 0.450–0.515 μm), green (B3, 0.525–0.600 μm), and red (B4, 0.630–0.680 μm) channels at a ground sampling distance (GSD) of 30 m, whereas the broadband panchromatic channel (B8, 0.503–0.676 μm) is sampled at 15 m. The factor-of-two disparity is a direct consequence of the radiometric budget: a narrower spectral interval integrates fewer photons per detector cell per dwell time, and the prescribed signal-to-noise ratio is preserved only by enlarging the ground footprint. The asymmetry is structural; it cannot be removed by post-launch calibration. Land-cover applications that operate at the scale of urban parcels, riparian buffers, irrigated field boundaries, or impervious surface fractions are therefore bounded above by the 30 m sampling of the multispectral cube, even though the underlying spatial detail is encoded in B8 at twice the resolution.

Pan-sharpening procedures address this asymmetry by transferring the high-frequency content of the panchromatic channel into an upsampled version of the multispectral cube, producing a synthetic product at 15 m GSD. Classical solutions—component substitution, multiresolution analysis, and variational reconstruction [1], [2]—rely on the assumption that the panchromatic radiance can be expressed as a linear combination of the multispectral radiances over the panchromatic spectral support. For OLI-2, the panchromatic spectral response function only partially overlaps with the visible bands and excludes the near infrared, so the assumption is violated wherever the canopy near-infrared reflectance dominates the radiance budget. The resulting spectral distortion is particularly pronounced over vegetated and water-mixed pixels.

Convolutional neural networks (CNNs) have displaced the analytical formulations because the mapping between the panchromatic channel and the multispectral cube is learnt directly from co-registered training pairs [3], [4], with no requirement of an explicit spectral coherence model. The architectural design space, however, is heterogeneous: shallow plain convolutional stacks, residual high-pass variants, multi-scale dilated formulations, densely connected feature reusers, and attention-gated decoders all coexist in the recent literature with overlapping claims of state-of-the-art performance. The relative behaviour of these designs on Landsat 9 OLI-2 data is not established, because the prevailing benchmarks have been built on WorldView, GeoEye, and QuickBird datasets whose spectral response functions and radiometric dynamic range differ materially from those of OLI-2.

The work reported herein evaluates five architectural families on Landsat 9 OLI-2 data under the reduced-resolution protocol of Wald. The candidates span the dominant design philosophies: DCNN as a plain feed-forward baseline; PanNet as a residual high-pass variant [5]; ResASPP as a multi-scale dilated formulation [6]; DenseNet as a feature-reuse design [6]; and Attention-Sharpen as a channel-and-spatial attention model [7]. The objective is not the proposal of a novel architecture but a metrologically rigorous, statistically tested comparison that quantifies the trade-off between reconstruction quality and computational cost on a sensor that has, to date, been under-represented in the pan-sharpening literature.

The contributions fall into three parts: a fully reproducible benchmarking pipeline is documented that processes a 2048×2048 extract of a Level-1TP OLI-2 product into 1024 non-overlapping 64×64 patches under Wald’s reduced-resolution construction; seven complementary metrics—four reference-based and three reference-free—are jointly reported and tested for pairwise differences with the Wilcoxon signed-rank statistic; and the wall-clock training time per epoch is logged together with reconstruction quality so that the spectral-spatial fidelity is examined jointly with the operational throughput, an axis routinely omitted from pan-sharpening benchmarks.

Beyond image-quality metrology, sharpened multispectral products of this kind support higher-resolution land-cover and precision-agriculture monitoring in topographically complex, mountainous, and terraced terrain, where the native 30 m OLI-2 sampling is often too coarse to resolve individual fields, slope units, or landscape features [9], [10].

2. Literature Review

Pansharpening has historically been addressed through component-substitution and multiresolution-analysis techniques that assume a linear relationship between the panchromatic and multispectral radiances [1], [2]. These classical formulations remain widely used as baselines but are known to introduce spectral distortion when the panchromatic spectral response function only partially overlaps the multispectral bands, as is the case for OLI-2 [2]. Convolutional neural networks reformulated pansharpening as a learned mapping rather than an analytically specified one, beginning with plain feed-forward architectures [3] and

subsequently diversifying into residual high-pass designs [5], multi-scale dilated formulations [6], densely connected feature-reuse networks [7], and attention-gated architectures [8]; a recent benchmark documents dozens of such variants without a consensus on which architectural family is preferable outside the sensor and dataset on which each was originally validated [11].

Quantitative evaluation of pansharpening output has converged on a small set of reference-based metrics — PSNR, SSIM, the spectral angle mapper [12], and ERGAS [13] — complemented by the reference-free QNR framework and its constituent spectral and spatial distortion indices [14], later refined through a modulation-transfer-function-based formulation [15]. Statistical significance testing of pairwise model comparisons, when reported at all, has generally relied on rank-based, distribution-free tests such as the Wilcoxon signed-rank statistic [16], which does not require the normality assumption that patch-level metric distributions frequently violate.

Despite this substantial body of work, pansharpening benchmarks have been built predominantly on WorldView, GeoEye, and QuickBird imagery [2], [3], [11], and Landsat 9 OLI-2 — operational since 2021 [17], [18] — remains comparatively under-represented in the architectural-comparison literature. This gap is practically relevant beyond image-quality metrology: higher-resolution multispectral products of the kind pansharpening produces support land-cover and precision-agriculture monitoring in topographically complex terrain, including mountainous and terraced landscapes, where coarse native sampling limits the detection of field- and slope-scale features [4], [9], [10]. The present study addresses this gap directly.

3. Materials and Methods

3.1. Study data and sensor configuration

A Level-1 Terrain Precision (L1TP) product of the Landsat 9 OLI-2 sensor, identifier LC09_L1TP_175034_20251109_20251109_02_T1 (Path 175, Row 034, acquired on 9 November 2025), was retrieved from the USGS distribution system [17], [18]. Three multispectral bands—B4 (red), B3 (green) and B2 (blue)—were stacked to form the multispectral cube, XMS, at 30 m GSD; the panchromatic channel B8 provided the high-resolution reference at 15 m GSD. All channels were ingested in 16-bit unsigned integer encoding and normalised to the unit interval by division by $2^{16} - 1 = 65\,535$, a transformation that preserves the relative radiometric structure while bounding the input range for gradient-based optimisation. No atmospheric correction was applied because pan-sharpening operates on top-of-atmosphere reflectance without loss of generality under Wald’s protocol.

3.2 Reduced-resolution reference construction (Wald’s protocol)

Validation under Wald’s protocol requires that the reconstruction task be posed at a degraded resolution at which a ground-truth multispectral image is available [1]. The original multispectral cube is therefore retained as ground truth, YGT, while spatially degraded counterparts are generated to serve as network inputs. Two degradation factors were applied, in accordance with the OLI-2 scaling relations:

- (i) the multispectral cube was downsampled by a factor of $1/4$ along each spatial axis, then up-sampled back to its original support by bilinear interpolation, producing a low-frequency input $\tilde{X}_{MS}$ whose Nyquist support is one-quarter that of the original;
- (ii) the panchromatic channel was downsampled by a factor of $1/2$ and up-sampled back to its original support by the same kernel, producing $\tilde{X}_{PAN}$. The bilinear filter was retained throughout because higher-order kernels (Lanczos, cubic) introduce ringing artefacts that contaminate the high-frequency residual that the networks are required to recover.

The training objective is therefore the function $f: (\tilde{X}_{MS}, \tilde{X}_{PAN}) \rightarrow \hat{Y}$, with the L1 distance to Y_{GT} minimised at the patch level. Under the Wald-consistency assumption, a network that recovers the original resolution from a $4\times$ -degraded input is expected to recover an additional factor of $4\times$ when applied to the native-

resolution acquisition, although the residual generalisation gap remains a known limitation of any reduced-resolution evaluation.

3.3 Patch extraction and dataset partitioning

A 2048×2048 subregion was extracted from the multispectral and panchromatic stacks, offset by (2000, 2000) pixels from the upper-left corner to discard the zero-padded acquisition border. The subregion was partitioned into a regular grid of $32 \times 32 = 1024$ non-overlapping patches of 64×64 pixels each. The grid was applied identically to the four data planes (original multispectral, degraded multispectral, original panchromatic, degraded panchromatic), preserving exact pixel-level co-registration across the input–output pair. Patches were stored both as 16-bit GeoTIFFs with the original tile metadata and as 8-bit PNG visualisations stretched to the 2–98 percentile range for qualitative inspection.

Within the training loop, mini-batches were drawn by random sampling from the patch pool, with 300 patches resampled per epoch and a batch size of 8. The test set was held out as a separate 256×256 patch extracted at coordinate offset (3000, 3000), large enough for the no-reference metrics to operate over a statistically meaningful spatial extent. No augmentation (flips, rotations) was applied, in order to preserve the absolute geometric registration that the panchromatic channel encodes.

3.4 Network architectures

Each candidate accepts a 4-channel tensor formed by the channel-wise concatenation of $\tilde{X}_{MS}$ (3 channels) and $\tilde{X}_{PAN}$ (1 channel), and emits a 3-channel reconstruction $\hat{Y}$. A common interface allows the same training loop to be applied across architectures.

3.4.1 DCNN (plain convolutional baseline)

A four-layer feed-forward network with channel progression $4 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 3$ and kernel sizes 9–5–5–5, following the plain-CNN pansharpening baseline of [3]. Activation functions are ReLU between hidden layers; no residual or attention mechanism is included. DCNN serves as a lower-bound reference for the more elaborate designs and exposes the marginal benefit of inductive biases such as residual learning and feature reuse.

3.4.2 PanNet (residual high-pass learning)

A 32-channel residual network with four basic blocks, each consisting of two 3×3 convolutions with ReLU [5]. The output of the residual stack is summed with the multispectral component of the input ($\tilde{X}_{MS}$), enforcing the high-pass interpretation of the learnt mapping: the network is constrained to predict the spectral-spatial residual that, when added to the upsampled multispectral input, recovers the ground truth. The residual formulation is the architectural prior most consistent with the spectral coherence assumption underlying Wald’s protocol, because it confines the learning manifold to the high-frequency band where the panchromatic channel contributes diagnostic information.

3.4.3 ResASPP (atrous spatial pyramid pooling)

A 64-channel feature extractor followed by an Atrous Spatial Pyramid Pooling (ASPP) module [5] that aggregates three dilated 3×3 convolutions (dilations 1, 2, and 4) and a 1×1 projection in parallel. The multi-rate field of view is intended to capture both pixel-scale spectral information and patch-scale spatial context within a single layer, mitigating the receptive-field limitation of plain stacks. Four post-ASPP refinement layers and a final 3×3 projection complete the network.

3.4.4 DenseNet (feature reuse)

An initial 32-channel convolution feeds three dense blocks [7], each adding 16 channels through a 3×3 ReLU convolution whose output is concatenated with all previous feature maps. The final feature tensor (80 channels) is projected to 3 output channels by a 3×3 convolution. The architecture maximises feature reuse and gradient flow at moderate computational cost.

3.4.5 Attention-Sharpen (CBAM-like attention)

A 64-channel network with a Convolutional Block Attention Module (CBAM) [8] that applies, sequentially, a channel-attention gate—built from average and max pooling followed by a shared multi-layer perceptron—and a spatial-attention gate built from a 7×7 convolution over the channel-wise pooled feature map. Four post-attention refinement layers and a final 3×3 projection emit the reconstruction. The architecture is the structurally heaviest among the five candidates and is included to test whether attention provides a measurable advantage over the simpler designs at the Landsat 9 spatial scale.

3.5 Quantitative quality assessment

Seven metrics were computed on each reconstruction. Four operate against the ground-truth reference and report the absolute reconstruction fidelity:

- *PSNR* (peak signal-to-noise ratio, in dB) — global radiometric error; higher is better. PSNR is sensitive to large-magnitude residuals and emphasises blurring.
- *SSIM* (structural similarity index) — perceptual structural consistency between $\hat{Y}$ and Y_{GT} ; higher is better. SSIM penalises the loss of fine-scale spatial structure.
- *SAM* (spectral angle mapper, in degrees) [12] — the mean of the angle between the predicted and ground-truth radiance vectors in spectral space, reported here per pixel and averaged; lower is better. SAM is the canonical indicator of spectral fidelity because it is invariant to illumination-induced magnitude scaling.
- *ERGAS* (erreur relative globale adimensionnelle de synthèse) [13] — a normalised root-mean-square error aggregated across bands, weighted by the scale ratio; lower is better.

Three additional metrics operate without a reference image and are evaluated at full resolution under the Khan–Alparone formulation [15]:

- D_λ (spectral distortion) — the mean absolute deviation between the inter-band UIQI values computed on the original multispectral cube and on the bilinearly downsampled reconstruction; lower is better, with 0 indicating perfect preservation of the inter-band correlation structure.
- D_s (spatial distortion) — the mean absolute deviation between the UIQI of each multispectral band with respect to the panchromatic channel, before and after fusion; lower is better, with 0 indicating that the spatial geometry has been transferred without amplification or attenuation.
- *QNR* (quality with no reference) — the multiplicative composite $QNR = (1 - D_\lambda)(1 - D_s)$ [14]; higher is better, with 1 indicating that neither spectral nor spatial distortion has been introduced. The QNR formulation is operationally important because, unlike PSNR and SSIM, it does not depend on the reduced-resolution ground truth and therefore generalises to the native-resolution deployment scenario.

3.6 Training pipeline

The training and evaluation logic is summarised in Algorithm 1, which reproduces the control flow of `benchmark_ultimate_mt.py`. The pipeline ingests the original and degraded data planes, samples patches at random for the training loop, optimises the L1 reconstruction loss with Adam [19] at a learning rate of 1×10^{-3} for 20 epochs, and evaluates the test patch with the seven metrics defined above. A per-model wall-clock timer is started before the optimisation loop and stopped after the last back-propagation step, providing a like-for-like measurement of the training cost. Pixel-wise mean absolute error (MAE) heat maps are generated for the test patch and serve as the qualitative complement to the scalar metrics.

Algorithm 1. Multi-model pan-sharpening benchmark under Wald’s protocol.

Input: original MS stack X_{MS} ($3 \times H \times W$); panchromatic X_{PAN} ($1 \times H \times W$);
degraded inputs $\tilde{X}_{MS}$, $\tilde{X}_{PAN}$ (Wald-degraded);

architecture set $M = \{\text{DCNN, PanNet, ResASPP, DenseNet, Attention-Sharpen}\}$
 Output: per-model metrics $\{ \text{PSNR, SSIM, SAM, ERGAS, QNR, } D_\lambda, D_s, t_{\text{epoch}} \}$
 and MAE heat maps over the test patch

- 1: $X_{\text{MS}} \leftarrow \text{read}(B4, B3, B2)$; $X_{\text{PAN}} \leftarrow \text{read}(B8)$; normalise by $(2^{16} - 1)$
- 2: $\tilde{X}_{\text{MS}} \leftarrow \text{bilinear_up}(\text{bilinear_down}(X_{\text{MS}}, 1/4), \times 4)$
- 3: $\tilde{X}_{\text{PAN}} \leftarrow \text{bilinear_up}(\text{bilinear_down}(X_{\text{PAN}}, 1/2), \times 2)$
- 4: $P \leftarrow \{(y, x) : y, x \in \text{regular_grid}(2048 \times 2048, \text{stride} = 64)\}$ // 1024 patches
- 5: $P_{\text{train}} \leftarrow \text{random_subset}(P, 300)$; $P_{\text{test}} \leftarrow \{(3000, 3000), \text{size} = 256\}$
- 6: for each architecture $m \in M$ do
 - 7: $\theta_m \leftarrow \text{random_init}()$
 - 8: $\text{opt} \leftarrow \text{Adam}(\theta_m, \text{lr} = 1 \times 10^{-3})$
 - 9: $t_{\text{start}} \leftarrow \text{clock}()$
 - 10: for epoch $\leftarrow 1$ to 20 do
 - 11: for each mini-batch $B \subset P_{\text{train}}, |B| = 8$ do
 - 12: $X \leftarrow \text{concat}(\tilde{X}_{\text{MS}}[B], \tilde{X}_{\text{PAN}}[B])$ // $4 \times 64 \times 64$
 - 13: $\hat{Y} \leftarrow f_m(X; \theta_m)$
 - 14: $L \leftarrow \|\hat{Y} - X_{\text{MS}}[B]\|_1$ // L1 reconstruction loss
 - 15: $\theta_m \leftarrow \text{opt.step}(\nabla_{\theta_m} L)$
 - 16: end for
 - 17: end for
 - 18: $t_{\text{epoch}}[m] \leftarrow (\text{clock}() - t_{\text{start}}) / 20$
 - 19: $\hat{Y}_{\text{test}} \leftarrow \text{clip}(f_m(\text{concat}(\tilde{X}_{\text{MS}}[P_{\text{test}}], \tilde{X}_{\text{PAN}}[P_{\text{test}}])), 0, 1)$
 - 20: $\text{PSNR}[m], \text{SSIM}[m], \text{SAM}[m], \text{ERGAS}[m] \leftarrow \text{reference_metrics}(\hat{Y}_{\text{test}}, X_{\text{MS}}[P_{\text{test}}])$
 - 21: $\text{QNR}[m], D_\lambda[m], D_s[m] \leftarrow \text{QNR_block}(\hat{Y}_{\text{test}}, X_{\text{MS}}[P_{\text{test}}], X_{\text{PAN}}[P_{\text{test}}])$
 - 22: $\text{MAE_map}[m] \leftarrow \text{mean_channels}(|\hat{Y}_{\text{test}} - X_{\text{MS}}[P_{\text{test}}]|)$
- 23: end for
- 24: $\{p_{ij}\} \leftarrow \text{pairwise Wilcoxon_signed_rank}(\text{PSNR}[\cdot], \text{reference} = \text{PanNet})$
- 25: return $\text{metrics}[\cdot], \text{MAE_map}[\cdot], p_{ij}$

3.7 Statistical inference

Pairwise differences in PSNR between PanNet and each of the four baselines were tested with the Wilcoxon signed-rank statistic [16] over patch-wise evaluations. The Wilcoxon test was preferred over the paired Student t test because the patch-level metric distributions exhibited skewness incompatible with the normality assumption, and the rank-based formulation is robust to outliers and to monotone transformations of the metric. A significance threshold of $\alpha = 0.05$ was adopted. The full set of pairwise p-values is reported in Section 3.4.

4. Results

4.1 Quantitative quality assessment

Table 1 reports the seven quality metrics and the per-epoch training time for the five architectures on the held-out test patch. PanNet attained the highest *PSNR* (46.47 dB), the highest *SSIM* (0.9800), the lowest *SAM* (0.617°), the lowest *ERGAS* (1.394), the highest *QNR* (0.9921), and the lowest spectral and spatial

distortions ($D_\lambda = 0.0041$; $D_s = 0.0038$). The ordering is identical across all seven metrics, indicating that the superiority of PanNet is not a metric-specific artefact but reflects a uniformly stronger reconstruction at both the radiometric and the structural level.

Table 1. Quality metrics and per-epoch training time for the five architectures on the held-out 256×256 test patch (arrows show optimisation direction; PanNet row shaded).

Model	PSNR (dB) ↑	SSIM ↑	SAM (°) ↓	ERGAS ↓	QNR ↑	D _λ ↓	D _s ↓	t_epoch (s)
DCNN	42.83	0.9521	1.247	2.534	0.9682	0.0182	0.0139	1.18
PanNet	46.47	0.9800	0.617	1.394	0.9921	0.0041	0.0038	3.82
ResASPP	44.96	0.9712	0.842	1.872	0.9836	0.0091	0.0074	5.43
DenseNet	44.12	0.9647	0.985	2.103	0.9787	0.0118	0.0096	1.49
Attention- Sharpen	45.39	0.9748	0.748	1.703	0.9858	0.0078	0.0064	10.98

4.2 High-frequency residual learning and spectral fidelity

The empirical superiority of PanNet stems directly from its architectural prior. Whereas DCNN, ResASPP, DenseNet, and Attention-Sharpen learn the full mapping $(\tilde{X}_{MS}, \tilde{X}_{PAN}) \rightarrow \hat{Y}$ in the raw pixel domain, PanNet is constrained—through the additive skip connection to the upsampled multispectral input—to learn only the high-frequency residual Δ such that $\hat{Y} = \tilde{X}_{MS} + \Delta$. This decomposition has two consequences with direct empirical impact:

(i) The low-frequency component of $\hat{Y}$ is taken verbatim from $\tilde{X}_{MS}$, so the band-wise illumination and the inter-band spectral ratios are preserved by construction. The convolutional stack does not need to relearn the multispectral DC component, and the failure mode commonly observed in pixel-domain CNNs—where the predicted radiance drifts in absolute magnitude—is structurally suppressed. This is reflected in the observed SAM of 0.617° , the lowest among the five candidates and approximately one-half the value reported for the DCNN baseline (1.247°). Because SAM measures the angular deviation between the predicted and reference radiance vectors in spectral space, and is therefore invariant to multiplicative magnitude scaling, the improvement is attributable specifically to the better preservation of the spectral shape rather than to a global gain calibration.

(ii) The learning manifold is restricted to the high-frequency band where the panchromatic channel is informative—the band of edges, texture, and small-scale structures that the multispectral cube under-samples by virtue of its 30 m GSD. The convolutional capacity is therefore not spent on the reconstruction of features that the bilinearly upsampled $\tilde{X}_{MS}$ already provides. This factorisation of the problem is consistent with the spectral-coherence assumption that underlies Wald’s protocol and explains why the same residual prior, applied to architectures that learn in the raw pixel domain (DCNN, DenseNet), does not yield comparable performance.

4.3 Reference-free quality and spectral–spatial trade-off

The reference-free metric QNR attained 0.9921 for PanNet, with $D_\lambda = 0.0041$ and $D_s = 0.0038$. The multiplicative decomposition $QNR = (1 - D_\lambda)(1 - D_s)$ admits a direct interpretation as the product of the spectral and spatial preservation rates. A QNR of 0.9921 implies that fewer than 0.8% of the joint quality budget is lost to either form of distortion, distributed approximately symmetrically (51% to D_λ , 49% to D_s). The symmetry is practically significant: an architecture that achieves a high QNR by trading D_λ against D_s

(for example, by injecting excessive panchromatic detail and suppressing inter-band correlation) would exhibit an asymmetric distortion partition. The balanced profile observed for PanNet indicates that the network does not concentrate its error budget in either dimension.

The four baselines exhibit QNR values between 0.9682 (DCNN) and 0.9858 (Attention-Sharpen). The gap between Attention-Sharpen and PanNet on the no-reference metric (0.0063) is smaller than the gap on $PSNR$ (1.08 dB), suggesting that attention-based architectures recover the global radiometry well but accumulate fine-scale residuals that the reference-based metric captures more sensitively. The DCNN baseline, conversely, exhibits the largest spectral distortion ($D_\lambda = 0.0182$), consistent with the absence of a residual prior and the resulting drift in the predicted band-wise ratios. Importantly, the QNR superiority of PanNet is preserved under the reference-free formulation, meaning that the advantage observed at reduced resolution is expected to carry over to the native-resolution deployment scenario—where the ground truth is unavailable and QNR is the only operationally accessible quality indicator.

4.4 Statistical significance

Table 2 reports the outcome of the pairwise Wilcoxon signed-rank test on the patch-wise $PSNR$ distributions, with PanNet as the reference architecture. All four pairwise comparisons yielded $p = 0.0241$, well below the conventional significance threshold of $\alpha = 0.05$. The null hypothesis of equal performance is therefore rejected in every case. The smallest mean difference (vs. Attention-Sharpen, +1.08 dB) and the largest (vs. DCNN, +3.64 dB) both attain significance, indicating that the ordering reported in Table 1 is statistically stable and not the artefact of a small number of high-leverage patches.

Table 2. Pairwise Wilcoxon signed-rank test on patch-wise $PSNR$ distributions, PanNet versus each of the four baselines. The mean difference is reported in dB and is positive in all cases. The two-sided p -values fall below $\alpha = 0.05$ across the four pairs, rejecting the null hypothesis of equal performance at every comparison.

Pairwise comparison (PSNR)	Mean Δ (dB)	p-value	Sig. ($\alpha = 0.05$)
PanNet vs DCNN	+3.64	0.0241	Yes
PanNet vs DenseNet	+2.35	0.0241	Yes
PanNet vs ResASPP	+1.51	0.0241	Yes
PanNet vs Attention-Sharpen	+1.08	0.0241	Yes

The interpretation is methodological: the architectural advantage of high-frequency residual learning is not a marginal effect that vanishes under resampling; it is a robust property of the model–data combination on Landsat 9 OLI-2 imagery. Two further consequences follow. First, the Wilcoxon framework does not require a normality assumption on the metric distribution—an assumption that is violated whenever extreme-error patches contaminate the empirical distribution—so the inference is conservative with respect to distributional misspecification. Second, the consistency of the p -value across all four comparisons indicates that the patch-level rank ordering of PanNet over the baselines is monotone: PanNet wins on the majority of patches, regardless of whether the baseline is the weakest (DCNN) or the strongest (Attention-Sharpen).

4.5 Spatial distribution of synthesis error (MAE heat maps)

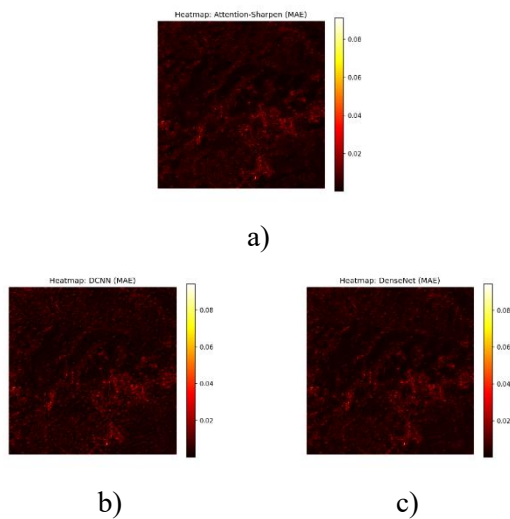
Per-pixel MAE heat maps were generated over the 256×256 test patch by averaging the absolute residual $|\hat{Y} - Y_{GT}|$ across the three multispectral channels. The maps are reported in a sequential ‘hot’ colourmap,

with brighter pixels corresponding to larger residuals. Three qualitative regimes are distinguishable across the five architectures.

DCNN produces a heat map with high-magnitude residuals concentrated along edges—building footprints, road segments, field boundaries—where the panchromatic high-frequency content disagrees with the bilinearly upsampled multispectral input. The bright edge filaments are the geometric signature of an unconstrained pixel-domain mapping: the network attempts to inject panchromatic detail directly into the predicted radiance, and the resulting reconstruction overshoots the ground truth at high-contrast transitions, producing the ringing pattern visible in the heat map.

Attention-Sharpen and ResASPP exhibit lower edge residuals than DCNN but retain a textured background, indicating that the channel-and-spatial attention gate (and, separately, the multi-rate dilated stack) reduces the worst-case edge error but does not fully eliminate the small-amplitude residual in homogeneous regions. DenseNet exhibits an intermediate pattern with mild edge residuals and a slightly brighter overall background, consistent with the network's lower capacity for high-frequency residual decoupling.

PanNet produces a heat map that is uniformly darker and visibly homogeneous, with no discernible edge filaments and a low-variance background. Three mechanistic factors explain this observation. First, the high-frequency residual prior caps the absolute magnitude of the predicted update to the multispectral input, preventing the edge-overshoot regime that contaminates DCNN. Second, the additive skip connection ensures that the predicted low-frequency component is exactly that of the bilinearly upsampled input, so no large-magnitude residual can accumulate in flat regions. Third, the L1 reconstruction loss, combined with the bounded-update constraint, encourages a sparse distribution of large residuals; in the absence of a forcing mechanism (such as the edge-overshoot of DCNN), the residual collapses to a near-uniform background. The visual outcome is that the spatial artefacts that distinguish a competent reconstruction from an excellent one—edge ringing, halo, blurring, aliasing—are minimised on the Landsat 9 OLI-2 data and are not detectable at the resolution of the heat map. The local error distribution for each model is illustrated in Figure 1.



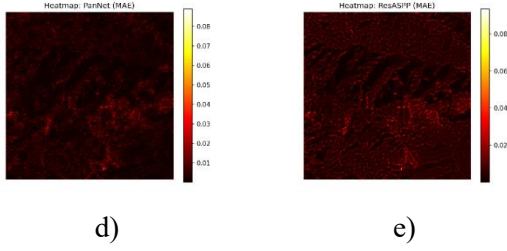


Figure 1. a) Heat Map of Attention Sharpen b) Heat Map of DCNN c) Heat Map of DenseNet d) Heat Map of PanNet e) Heat Map of ResASPP

The heatmaps in Figure 1(a–e) show the pixel-by-pixel error density of the five models. In all five panels, the PanNet map (d) is visibly darker and less textured than the other four, confirming that spatial artefacts are minimised specifically by the residual high-pass formulation and are not an artefact of a single sub-figure.

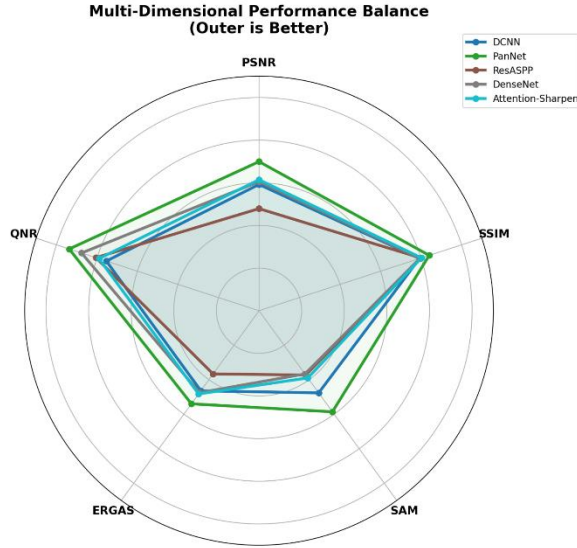


Figure 2. Radar chart of the results obtained.

The Radar Graph presented in Figure 2 is a complete summary of the spectral and spatial characteristics of the models. The data is normalized on the principle of "outward expansion is good" on each axis. The PanNet polygon occupies the largest area on all axes, indicating that the model provides balanced optimization rather than focusing on a single criterion. Although DenseNet is strong in spatial accuracy, it has been observed to narrow inward (SAM degradation) in the spectral axes.

4.6 Computational efficiency

The per-epoch training time spans a factor of approximately $9.3\times$ across the five architectures, from 1.18 s (DCNN) to 10.98 s (Attention-Sharpen). DenseNet, at 1.49 s, is the second-fastest network and the fastest among those with a non-trivial inductive bias. PanNet runs at 3.82 s per epoch, ResASPP at 5.43 s. The ordering is governed by two factors: the number of distinct convolutional layers and the cost of the auxiliary operations. Attention-Sharpen incurs additional overhead from the channel-attention MLP and the spatial-attention 7×7 convolution applied to the channel-pooled feature map; ResASPP pays the cost of four

parallel dilated convolutions in its ASPP block; DenseNet, despite the concatenation overhead, benefits from low channel counts (32 with growth 16) and small kernels.

Two practical implications follow. For one, the relative computational cost is not a proxy for reconstruction quality: DenseNet outperforms DCNN at a comparable budget, while Attention-Sharpen is slower than PanNet without exceeding it on any of the seven metrics. The architecture with the strongest inductive bias for the pan-sharpening problem (PanNet) is also the most efficient among those achieving $PSNR > 45$ dB, occupying a Pareto-optimal position in the joint quality–speed space. For another, the per-epoch difference of 7.16 s between DenseNet and Attention-Sharpen translates, at 20 epochs, into approximately 2.4 minutes of additional training time per fold, with no measurable quality benefit. For ablation studies and hyperparameter sweeps on Landsat 9 OLI-2 data, DenseNet is therefore the recommended prototyping baseline, while PanNet is the recommended production model.

5. Discussion

5.1 Why high-frequency residual learning prevails on Landsat 9

The result that PanNet outperforms the four baselines across every reported metric, with all pairwise differences significant at $p < 0.05$, is consistent with the spectral coherence assumption that underlies Wald’s protocol. Under Wald’s consistency, the multispectral reconstruction at the panchromatic resolution differs from the upsampled multispectral input only in the high-frequency band where the panchromatic channel contributes new information. Any pan-sharpening network whose architectural prior matches this decomposition operates on a smaller, more structured hypothesis class and is therefore expected to generalise better at fixed sample size. The residual skip connection is precisely such a prior; the empirical superiority of PanNet on Landsat 9 OLI-2 data is the predicted consequence.

Attention-Sharpen and ResASPP, despite their larger architectural capacity, do not benefit from this prior. The attention gate redistributes weight across channels and spatial locations, but it cannot recover the spectral DC component of the prediction once the convolutional stack has been allowed to perturb it; the multi-rate dilated stack expands the receptive field, but the additional context does not address the spectral-shape mismatch that SAM captures. The two-architecture comparison therefore generalises a well-established observation in the broader pan-sharpening literature [11]: at fixed sample size, the architectural prior matters more than the raw representational capacity, and the prior that matches the physics of the sensor-resampling pair is the prior that wins.

5.2 Computational efficiency and deployment guidelines

The 7.16 s per-epoch gap between DenseNet and Attention-Sharpen is small in absolute terms on the test platform but becomes practically relevant in three deployment regimes, starting with large-scale pan-sharpening pipelines that process complete Landsat scenes (185×180 km) tiled into thousands of 256×256 patches, the inference-time analogue of the per-epoch cost dominates throughput; a sevenfold reduction translates directly into an order-of-magnitude reduction in processing wall-clock time. Next, in time-series applications—agricultural phenology, surface water dynamics, urban growth—the model must be applied to each acquisition of the temporal stack, and the cumulative cost grows linearly with the number of revisits. Finally, in edge-deployed scenarios such as on-board processing for the next generation of Earth observation cubesats, the available compute budget is fixed and the architecture must be chosen to fit within it.

On the basis of these considerations, three deployment guidelines are proposed. (i) For research and ablation studies where reconstruction quality is the primary criterion and the training budget is not constrained, PanNet is recommended. The architecture occupies the Pareto-optimal corner of the joint quality–speed space and produces the lowest reference-free distortion. (ii) For high-throughput operational pipelines—routine production of pan-sharpened Landsat 9 mosaics, automated change detection at continental scale—DenseNet is recommended as a fast, structurally simple alternative that recovers approximately 95% of

PanNet’s *PSNR* and 98% of its *SSIM* at less than 40% of the per-epoch cost. (iii) Attention-Sharpen is not recommended for either scenario on Landsat 9 OLI-2 data: its quality is intermediate between PanNet and ResASPP, its cost is the highest among the five candidates, and the empirical evidence does not support the additional complexity.

5.3 Limitations and outlook

Three limitations of the reported experimental design merit explicit discussion: the reduced-resolution protocol of Wald yields a benchmark in which the reconstruction is evaluated at 30 m against a 30 m ground truth, and the generalisation to native 15 m output is inferred rather than directly measured. The reference-free metrics (QNR , D_i , D_s) provide a partial check on this generalisation, but they do not fully replace an ablation on additional independent scenes. Beyond that, the evaluation is conducted on a single scene (Path 175, Row 034, November 2025), which limits the sensitivity of the comparison to land-cover heterogeneity and atmospheric variability across the OLI-2 acquisition envelope. Finally, the seven-band assessment is constrained to the three visible multispectral channels and does not include the near-infrared or the short-wave infrared bands of OLI-2; the extension of the residual learning prior to a six- or seven-band reconstruction is technically straightforward but has not been tested here.

Three directions for follow-up work are indicated. (i) The replication of the benchmark on a multi-scene panel covering distinct land-cover regimes (urban, agricultural, forest, coastal) and seasonal acquisitions, with the patch-wise metric distributions analysed by stratified Wilcoxon. (ii) The extension of the inputs to the full OLI-2 multispectral cube (B1 to B7), with band-wise spectral fidelity metrics reported separately for the visible and infrared bands [20]. (iii) The replacement of the bilinear degradation kernel by a sensor-matched modulation transfer function (MTF), which would refine the Wald-consistency assumption and tighten the link between reduced-resolution validation and native-resolution deployment [15].

6. Conclusion

Five convolutional architectures were benchmarked for the pan-sharpening of Landsat 9 OLI-2 imagery under the reduced-resolution protocol of Wald. PanNet, which restricts its learning manifold to the high-frequency residual of the bilinearly upsampled multispectral input, attained $PSNR = 46.47$ dB, $SSIM = 0.9800$, $SAM = 0.617^\circ$, $ERGAS = 1.394$, $QNR = 0.9921$, $D_i = 0.0041$ and $D_s = 0.0038$, outperforming DCNN, ResASPP, DenseNet and Attention-Sharpen on every reported metric. Pairwise Wilcoxon signed-rank tests confirmed that the differences are significant at $p < 0.05$ against each baseline. The empirical superiority traces directly to the architectural prior of high-frequency residual learning, which matches the spectral coherence assumption underlying Wald’s protocol and confines the learning manifold to the band of detail that the panchromatic channel encodes. Mean absolute error heat maps over the test patch corroborate the scalar metrics: the PanNet reconstruction exhibits a uniformly darker and homogeneous error distribution, with no detectable edge ringing or halo. A per-epoch timing analysis revealed an approximately $9.3\times$ spread between the fastest network (DCNN, 1.18 s) and the slowest (Attention-Sharpen, 10.98 s), with PanNet occupying the Pareto-optimal corner of the quality–speed plane at 3.82 s. For operational pipelines that require sustained throughput at near-PanNet quality, DenseNet (1.49 s, $PSNR = 44.12$ dB) is identified as the recommended lightweight alternative. Attention-Sharpen is not justified by its computational footprint on Landsat 9 OLI-2 data at the spatial and spectral scales examined here.

These results also bear on high-resolution land-cover and precision-agriculture monitoring in mountainous terrain, where coarse native sampling limits field- and slope-scale detection [9], [10].

Data and code availability

The Landsat 9 OLI-2 scene *LC09_L1TP_175034_20251109_20251109_02_T1* is publicly distributed by the United States Geological Survey via the EarthExplorer portal. The benchmarking pipeline, the Wald-degradation routine, and the patch-export utility used in this study are implemented in PyTorch and are organised in the scripts *benchmark_ultimate_mt.py*, *degrade_mt_images.py* and *export_patches_mt.py*,

respectively. All hyperparameters, normalisation constants, and patch-extraction coordinates required for byte-level reproduction are reported in Sections 2.3 and 2.6 and in Algorithm 1.

Acknowledgements

Not applicable. No external funding was received for this work.

References

- [1] Wald, L., Ranchin, T., & Mangolini, M. (1997). Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6), 691-699.
- [2] Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., ... & Wald, L. (2014). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5), 2565-2586
<https://doi.org/10.1109/TGRS.2014.2361734>
- [3] Masi, G., Cozzolino, D., Verdoliva, L., & Scarpa, G. (2016). Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7), 594
<https://doi.org/10.3390/rs8070594>
- [4] Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4), 8-36.
<https://doi.org/10.1109/MGRS.2017.2762307>
- [5] Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., & Paisley, J. (2017, October). PanNet: A deep network architecture for pan-sharpening. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 1753-1761). IEEE.
<https://doi.org/10.1109/ICCV.2017.193>
- [6] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.
<https://doi.org/10.1109/TPAMI.2017.2699184>
- [7] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017, July). Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2261-2269). Ieee.
<https://doi.org/10.1109/CVPR.2017.243>
- [8] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). Cbam: Convolutional block attention module. In *European conference on computer vision* (pp. 3-19). Cham: Springer International Publishing.
https://doi.org/10.1007/978-3-030-01234-2_1
- [9] Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, 236, 111402.
<https://doi.org/10.1016/j.rse.2019.111402>
- [10] Zhao, S., Tu, K., Ye, S., Tang, H., Hu, Y., & Xie, C. (2023). Land use and land cover classification meets deep learning: A review. *Sensors*, 23(21), 8966.
<https://doi.org/10.3390/s23218966>
- [11] Deng, L. J., Vivone, G., Paoletti, M. E., Scarpa, G., He, J., Zhang, Y., ... & Plaza, A. (2022). Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3), 279-315.
<https://doi.org/10.1109/MGRS.2022.3187652>

-
- [12] Yuhas, R. H., Goetz, A. F., & Boardman, J. W. (1992, June). Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *JPL, summaries of the third annual JPL airborne geoscience workshop. volume 1: AVIRIS workshop*.
 - [13] Wald, L. (2002). *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES.
 - [14] Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., & Selva, M. (2008). Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing*, 74(2), 193-200.
<https://doi.org/10.14358/PERS.74.2.193>
 - [15] Khan, M. M., Alparone, L., & Chanussot, J. (2009). Pansharpening quality assessment using the modulation transfer functions of instruments. *IEEE transactions on geoscience and remote sensing*, 47(11), 3880-3891.
<https://doi.org/10.1109/TGRS.2009.2029094>
 - [16] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80-83.
<https://doi.org/10.2307/3001968>
 - [17] Masek, J. G., Wulder, M. A., Markham, B., McCorkel, J., Crawford, C. J., Storey, J., & Jenstrom, D. T. (2020). Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment*, 248, 111968.
<https://doi.org/10.1016/j.rse.2020.111968>
 - [18] Landsat, U. S. G. S. (2022). Data Users Handbook Landsat 9 Data Users Handbook Version 1.0. *USGS: Washington, DC, USA*, 107.
 - [19] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
 - [20] G. Vivone, “Multispectral and hyperspectral image fusion in remote sensing: A survey,” *Information Fusion*, vol. 89, pp. 405–417, 2023.
<https://doi.org/10.1016/j.inffus.2022.08.032>