

Radiometrically Constrained Multi-Temporal GAN Super-Resolution of Sentinel-2 Imagery over India

Aveek Sarkar

Abstract—The 10 m spatial resolution of the Sentinel-2 visible bands limits the delineation of narrow roads, small buildings and smallholder field boundaries. Generative super-resolution (SR) can recover part of this structure, but existing models are trained predominantly on imagery from the United States, and their outputs adopt the radiometry of the high-resolution training reference instead of the measured Sentinel-2 reflectance. This paper addresses both limitations for the Indian subcontinent. The Satlas eight-date ESRGAN is fine-tuned on 103,936 Sentinel-2 L2A image pairs from 55 sites, using ArcGIS World Imagery at 2.39 m as the reference (`arcgis_B`). A second model, `s2colour`, embeds a parameter-free band-value constraint in the generator and is trained with colour-invariant losses and a high-pass discriminator, so that the mean B04, B03 and B02 values of every 10 m cell equal those of the input. On 2,800 tiles from seven held-out sites covering urban, paddy, desert, forest, hill and snow terrain, `arcgis_B` improves edge-F1 from 0.524 to 0.642 and LPIPS from 0.382 to 0.182 relative to the original Satlas model, and both proposed models outperform SEN2SR, LDSR-S2 and Real-ESRGAN on these metrics at all seven sites. `s2colour` attains comparable structural quality (edge-F1 0.638, LPIPS 0.193) while reducing the reflectance error with respect to the input from 0.095 to 0.022. Applying the same constraint to `arcgis_B` after training reduces edge-F1 to 0.528 (paired difference +0.110, 95% CI [0.107, 0.112]). Against same-day VEN μ S surface reflectance, the band error of `s2colour` exceeds the inter-sensor discrepancy by 0.004. The study further derives an upper bound of 3.01 dB on the PSNR penalty of a posterior-sampling estimator, reports the geometric consistency of the multi-date input, and validates a per-pixel confidence layer against the observed error (AUROC 0.743). Code, model weights and all evaluation outputs are publicly released.

Index Terms—Super-resolution, Sentinel-2, generative adversarial networks, multi-temporal fusion, radiometric consistency, domain adaptation, uncertainty estimation.

I. INTRODUCTION

THE Sentinel-2 constellation provides systematic global coverage with a revisit interval of five days and an archive extending to 2015, at a spatial resolution of 10 m in the visible bands [1]. This resolution is insufficient for several mapping tasks of particular relevance to India, where historic urban cores consist of blocks a few metres wide and many agricultural parcels are narrower than two Sentinel-2 pixels. Commercial very-high-resolution (VHR) imagery resolves such features, but its cost and acquisition frequency restrict its use for routine monitoring and for time-critical applications such as flood and disaster assessment.

Super-resolution offers a means of increasing the effective resolution of the freely available Sentinel-2 record. Networks optimised for pixel-wise losses converge towards the conditional mean of plausible high-resolution scenes, which is spatially smooth, whereas adversarially trained networks produce

sharper reconstructions [2], [3], [4]. In the satellite setting, multiple acquisitions of the same area provide additional information, since sub-pixel offsets between revisits sample the scene beyond the Nyquist limit of a single image [5], [6], [7]. The Satlas model [8] combines both principles in an ESRGAN that ingests eight Sentinel-2 acquisitions and was trained on 1.2 million Sentinel-2/NAIP pairs from the United States.

Two limitations restrict the use of such models over India. The first is a domain shift: urban morphology, roofing materials, field patterns and vegetation differ substantially from those of the training region. On the Indian test set used in this study, the released Satlas model exhibits the highest opensr-test hallucination score (0.372) and the highest LPIPS (0.382) among the generative models evaluated. The second is radiometric. A generative model reproduces the colour characteristics of its training reference; when the reference is an aerial or VHR basemap, the output no longer corresponds to Sentinel-2 surface reflectance. This precludes comparison with other Sentinel-2 products, the computation of spectral indices and quantitative change detection. Correcting the colour after inference is shown in Section VI-D to remove a large part of the reconstructed detail.

This paper presents two models that address these limitations (Fig. 1). `arcgis_B` is a domain-adapted version of Satlas intended for visual interpretation and digitisation. `s2colour` shares its architecture and initialisation but constrains the output to the measured Sentinel-2 band values at 10 m, and is intended for quantitative analysis. Both models are evaluated on seven geographically separate sites against openly available Sentinel-2 SR models and with the opensr-test benchmark [9]. The contributions are as follows:

- 1) A band-value constraint embedded in the generator, which fixes the B04, B03 and B02 mean of every 10 m cell to the observed value, together with a training scheme (colour-invariant losses and a high-pass discriminator) that allows the network to be optimised through the constraint. A paired bootstrap analysis over 2,800 tiles shows that applying the same constraint after training yields significantly lower structural fidelity.
- 2) A multi-temporal generative SR model adapted to India, trained on 103,936 Sentinel-2 L2A / 2.39 m pairs from 55 sites with a complete metadata audit, which outperforms the published models tested on structural and perceptual metrics at all seven held-out sites.
- 3) An independent assessment of radiometric and geometric consistency, comprising a same-day comparison with VEN μ S surface reflectance [10], the co-registration accuracy of the input time series and the positional



Fig. 1. Hyderabad old city (held-out test site). Columns: ArcGIS World Imagery reference at 2.39 m; one Sentinel-2 acquisition at 10 m (input); `arcgis_B` (proposed); `s2colour` (proposed, Sentinel-2 band values constrained per 10 m cell); ESA SEN2SR Lite; ESA LDSR-S2; Real-ESRGAN. Reference imagery © Esri and its providers, reproduced for evaluation purposes only.

accuracy of the output.

- 4) A quantitative treatment of the perception–distortion trade-off for this task, including a bound on the PSNR penalty of a posterior-sampling estimator, a decomposition of the added detail into correct and hallucinated components, and a per-pixel confidence layer validated against the observed error.

II. RELATED WORK

Generative super-resolution. SRGAN [2] introduced adversarial and perceptual losses for SR and demonstrated that PSNR and SSIM do not reflect perceived quality. ESRGAN [3] introduced residual-in-residual dense blocks, a relativistic discriminator and pre-activation perceptual features. Blau and Michaeli [4] proved that distortion and perceptual quality cannot be optimised simultaneously. Real-ESRGAN [11] is trained on synthetic degradations with a spectrally normalised U-Net discriminator. Transformer architectures such as SwinIR [13] and, for remote sensing, Swin2-MoSE [14] improve distortion metrics in the single-image setting.

Multi-image super-resolution in Earth observation. The PROBA-V challenge [5] established multi-image SR for satellite data and introduced the shift-tolerant cPSNR metric. HighRes-net [6] fuses frames recursively, and WorldStrat [7] reports consistent gains from four to eight revisits. Satlas [8] extended the approach to an eight-date ESRGAN and reported that PSNR and SSIM correlate poorly with human judgement and that diffusion models produce realistic but inaccurate outputs. Retnanto et al. [15] combine single- and multi-image SR, and Kowaleczko et al. [16] observe in a real-world benchmark that the highest PSNR and SSIM are obtained by blurred outputs.

Radiometric consistency and evaluation. SEN2SR [17] includes a low-frequency hard-constraint layer so that the output, when degraded to 10 m, reproduces the input. LDSR-S2 [18] is a latent diffusion model that provides a per-pixel

TABLE I
PROPERTIES OF THE OPENLY AVAILABLE SENTINEL-2 SR MODELS COMPARED IN THIS STUDY, AS DESCRIBED IN THE CITED WORKS.

Model	dates	training data	band constraint	confidence
Real-ESRGAN [11]	1	synthetic	–	–
SEN2SR [17]	1	US	low-freq. layer	–
LDSR-S2 [18]	1	US	–	✓
Satlas [8]	8	US (NAIP)	–	–
<code>arcgis_B</code> (proposed)	8	India	–	✓
<code>s2colour</code> (proposed)	8	India	in-generator	✓

uncertainty estimate. Both are single-image models trained on imagery from the United States. The opensr-test framework [9] separates consistency with the input from correctness with respect to a reference, and decomposes the added detail into improvement, omission and hallucination. SEN2VEN μ S [10] provides Sentinel-2 acquisitions paired with same-day VEN μ S reflectance at 5 m.

Table I summarises the properties of the compared models. To the author’s knowledge, no previous model combines multi-temporal input, a hard band-value constraint and a validated per-pixel confidence output. The present study further finds that such a constraint must be part of the generator during training if the level of detail characteristic of adversarial models is to be retained.

III. METHOD

A. Problem formulation

The input y is a stack of eight co-located Sentinel-2 L2A acquisitions (bands B04, B03 and B02) resampled to the zoom-17 Web Mercator grid at 9.55 m/px, with 32×32 px per tile. The target is the corresponding tile at 2.39 m/px (128×128 px). The two grids nest exactly at a factor of four, so that each input cell corresponds to a 4×4 block of output pixels and the output can be exported as georeferenced tiles

without resampling. The training reference consists of ArcGIS World Imagery tiles on the same grid (Appendix A).

B. Generator

Both models use the Satlas generator, an ESRGAN RRDB-Net [3] in which the first convolution receives the eight acquisitions stacked as 24 channels. It is followed by 23 residual-in-residual dense blocks, a global skip connection, two $\times 2$ upsampling stages and two output convolutions, for a total of 16.7M parameters. Stacking the acquisitions along the channel dimension allows the first layer to distinguish persistent structure from transient effects such as cloud, haze or land-cover change, and to exploit the sub-pixel offsets between dates.

C. Adapting Satlas to India (*arcgis_B*)

The released Satlas eight-date weights are fine-tuned on Indian image pairs using the loss configuration of Real-ESRGAN [11]: L1 and VGG19 perceptual losses computed against an unsharp-masked target, and an adversarial loss from a spectrally normalised U-Net discriminator. Three modifications are introduced to limit hallucinated detail:

- the discriminator is conditioned on the eight upsampled input dates, so that it assesses realism with respect to the observed scene and penalises structure that is inconsistent with the input;
- the adversarial loss weight is reduced from 0.1 to 0.05;
- an edge loss, defined as the L1 distance between the Sobel gradients of output and target (weight 0.5), is added; it is insensitive to colour and to homogeneous regions.

The unmodified Satlas configuration performed worse on five of six validation metrics. The released weights are the exponential moving average of the generator parameters. Training settings are listed in Appendix B.

D. Band-value constraint

Let A denote the 4×4 block average (one 10m cell), U bicubic $\times 4$ upsampling, $\hat{x}$ the output of the generator body and m the per-pixel median of the eight input dates, each applied per band. The constraint applies the update

$$\hat{x} \leftarrow \hat{x} + U(m - A\hat{x}) \quad (1)$$

three times. Each iteration reduces the difference between the block means of $\hat{x}$ and m through a smooth correction. Since AU approximates but does not equal the identity, the update is iterated; nearest-neighbour upsampling would satisfy the constraint in a single step but introduces visible discontinuities at block boundaries. The operator has no learnable parameters and can therefore be applied to any generator checkpoint.

Proposition 1. *If $A\hat{x} = m$, then for any set Ω of complete 10m cells and any linear functional ℓ of the cell means (for example a band mean over Ω , a zonal sum or a band difference), $\ell(\hat{x}) = \ell(m)$. Moreover, the residual $\hat{x} - U'(m)$ has zero mean in every cell for any block-mean-preserving upsampling operator U' .*

Proof. Since averaging is linear, the mean of $\hat{x}$ over Ω is the mean of its block means, which equal those of m ; the same argument applies to any linear combination of cell means. For the second statement, $A(\hat{x} - U'(m)) = m - m = 0$. $\square$

The proposition implies that the generated high-frequency content redistributes the value of each 10m cell among its 16 output pixels without altering the cell mean. Statistics aggregated over complete 10m cells are therefore identical to those of the Sentinel-2 input, up to the residual of the iterative update, which is quantified in Appendix D.

E. Training with the constraint (*s2colour*)

s2colour uses the same generator with (1) as its final operation. Each reference tile exhibits a colour offset that cannot be predicted from the input (the mean absolute per-channel bias on validation tiles is 7/255, of which 83% varies between tiles). The training objective is therefore designed to be insensitive to this offset:

- *Colour-invariant losses.* The L1 and perceptual targets are shifted to the per-channel mean of the output, $\tilde{y} = y + \text{sg}(\mu(\hat{x}) - \mu(y))$, where sg denotes the stop-gradient operator, so that these losses measure structural differences only.
- *High-pass discriminator.* The discriminator receives $\hat{x} - U(A\hat{x})$, the component of the output at scales finer than one input cell, together with the eight input dates, and thus evaluates texture rather than colour.

Under this scheme, the colour of the output is determined by the input through the constraint, and the training signal acts on spatial detail only. *s2colour* is initialised from the *arcgis_B* generator and a newly initialised high-pass discriminator.

F. Per-pixel confidence

As an indicator of where the output is inferred rather than observed, the magnitude of the detail added to a plain interpolation of the input is used, $c = |\hat{x} - U_b(m)|$ averaged over bands, where U_b denotes bicubic upsampling. The layer requires a single interpolation and is validated against the observed error in Section VI-H.

IV. PERCEPTION-DISTORTION ANALYSIS

Since a model that produces sharp reconstructions is expected to score lower on PSNR, the magnitude of this penalty is derived here. Let y denote the input and x the true 2.39m scene, with $\mu = \mathbb{E}[x | y]$ and $V = \mathbb{E}[\|x - \mu\|^2 | y]$.

Minimum mean squared error estimator. For any estimator $\hat{x}$ that depends only on y ,

$$\mathbb{E}[\|\hat{x} - x\|^2 | y] = \|\hat{x} - \mu\|^2 + V, \quad (2)$$

which is minimised by $\hat{x} = \mu$, the pixel-wise average over all plausible scenes. Where the position of an edge is uncertain, this average yields a gradual transition rather than an edge.

Posterior sampling. If $\hat{x} \sim p(x | y)$ is drawn independently of x given y , as an ideal generative model would do, then

$$\mathbb{E}[\|\hat{x} - x\|^2 | y] = \mathbb{E}\|\hat{x} - \mu\|^2 + \mathbb{E}\|x - \mu\|^2 = 2V. \quad (3)$$

The PSNR of the posterior sample is therefore lower by $10 \log_{10} 2 \approx 3.01$ dB, while its output follows the distribution of real scenes. This is the squared-error case of the general result of Blau and Michaeli [4].

Example. Consider a road edge located at one of four equally likely positions within one 10m pixel. The posterior mean is the ramp (0.25, 0.50, 0.75, 1.00), with an expected squared error of 0.625 and no detectable edge. A sharp edge placed at a randomly drawn plausible position has an expected error of $1.25 = 2 \times 0.625$, but represents the road within two output pixels of its true position and inside the 10m pixel constrained by the input.

The penalty measured in this study is 0.19–0.52 dB PSNR relative to bicubic interpolation, well below the bound, with corresponding gains of +0.23 in edge-F1 and −0.39 in LPIPS (Section VI). In line with the recommendations of the Satlas [8], Kowaleczko et al. [16] and opensr-test [9] authors, models are ranked by structural and perceptual metrics, and pixel-wise metrics are reported for completeness.

V. EXPERIMENTAL SETUP

A. Data

The training set comprises 103,936 image pairs from 55 sites in India, covering 23 cities of differing morphology, 11 agricultural regions (wheat, paddy, chilli, soybean and sugarcane), desert, scrubland, cold desert, forest, tea plantations, coastal areas and Himalayan snow cover (Fig. 7). The input consists of Copernicus Sentinel-2 L2A data. The reference consists of ArcGIS World Imagery tiles at 2.39 m, resampled from 0.31–0.5 m WorldView-2/3, GeoEye-1 and Legion acquisitions. For each tile, eight cloud-screened Sentinel-2 acquisitions within ± 120 days of the reference capture date are selected (median offset 20 days). The acquisition date, sensor and native resolution of every reference tile were verified from the imagery metadata. The construction, filtering and audit procedures are described in Appendix A.

B. Test protocol

Seven sites with 400 tiles each (2,800 tiles in total) were collected separately and excluded from training and checkpoint selection: Hyderabad old city and Varanasi (urban), Thanjavur (paddy), Barmer (desert), Wayanad (forest), Shillong (hills) and Lachung, North Sikkim (snow). Every test tile lies at least 54.3 km from the nearest training tile (76–223 km for all sites except Shillong). Six leakage checks, covering shared tile identifiers, byte-identical reference images and overlap with the training regions of either initial model, returned no matches. A second version of the test set uses Sentinel-2 acquisitions approximately six months from the reference capture date (off-season), when crop, water and snow conditions differ.

C. Evaluation metrics

A 4px border is excluded from all tile-level metrics.

- *Edge-F1* (primary structural metric): F1 score between Canny edge maps of contrast-normalised, Gaussian-smoothed ($\sigma = 2$ px) greyscale images, with a matching tolerance of 1px; the chance level is 0.38.

- *Gradient correlation*: Pearson correlation between Sobel gradient magnitudes.
- *LPIPS* [12] (AlexNet backbone): primary perceptual metric.
- *PSNR*, *SSIM* [19] and *cPSNR* [5].
- *opensr-test* [9] (v1.3.3), computed on 4×4 -tile chips: improvement, omission and hallucination, which partition the added detail and sum to one, and the reflectance, spectral and spatial consistency with the input.

Since the colour of the ArcGIS reference is a product-specific rendering, PSNR, SSIM and LPIPS are computed after matching the per-channel mean and standard deviation of each output to the reference. Edge-F1 and gradient correlation are independent of colour by construction.

D. Baseline models

All published models are run according to their reference implementations, including band order, scaling, tiling and clamping. SEN2SR Lite [17] and LDSR-S2 [18] receive bands B02, B03, B04 and B08 of the acquisition closest to the reference capture date, processed as 4×4 -tile chips. Real-ESRGAN [11] (general x4v3) receives the same acquisition as a true-colour composite. The original Satlas model [8] receives the same eight acquisitions as the proposed models. Bicubic interpolation is reported for a single date and for the eight-date median. The choice of a multi-temporal GAN over transformer and diffusion architectures was based on an equal-budget comparison of 18 models (Appendix C).

VI. RESULTS

A. Comparison with published models

Table II reports the in-season results. Fine-tuning on Indian data improves all structural and perceptual metrics of Satlas: edge-F1 increases from 0.524 to 0.642, gradient correlation from 0.139 to 0.356 and the opensr-test improvement score from 0.423 to 0.510, while LPIPS decreases from 0.382 to 0.182 and the hallucination score from 0.372 to 0.264. Relative to LDSR-S2, the strongest single-image baseline, *arcgis_B* achieves an edge-F1 higher by 0.15 and an LPIPS approximately half as large.

The single-image models add comparatively little high-frequency content, and their omission and hallucination scores are close to those of bicubic interpolation. The proposed models add the largest share of correct detail, at the cost of a higher hallucination score than the single-image models (Fig. 2). On pixel-wise metrics the ranking is reversed, consistent with Section IV: SEN2SR obtains the highest PSNR, and bicubic interpolation exceeds both proposed models in cPSNR (20.78 compared with 20.39–20.41).

B. Results by land-cover type

Table III reports the results per site. One of the proposed models obtains the highest edge-F1 and the lowest LPIPS at each of the seven sites. The smallest margin occurs in the closed forest of Wayanad, where LDSR-S2 (0.536) is within 0.001 of *arcgis_B* and below *s2colour* (0.554); dense

TABLE II

RESULTS ON THE SEVEN HELD-OUT INDIAN SITES, IN-SEASON (2,800 TILES; OPENSr-TEST ON 117 CHIPS). $\uparrow/\downarrow$: HIGHER/LOWER IS BETTER. PSNR, SSIM AND LPIPS ARE COLOUR-NORMALISED. “BAND ERROR” DENOTES THE OPENSr-TEST REFLECTANCE CONSISTENCY WITH THE INPUT. BEST VALUES IN BOLD.

Model	dates	edge-F1 $\uparrow$	grad-corr $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	cPSNR $\uparrow$	improv. $\uparrow$	omission $\downarrow$	halluc. $\downarrow$	band error $\downarrow$
arcgis_B (proposed)	8	0.642	0.356	0.182	0.473	20.15	20.39	0.510	0.226	0.264	0.095
s2colour (proposed)	8	0.638	0.344	0.193	0.446	19.82	20.41	0.472	0.288	0.239	0.022
Satlas [8]	8	0.524	0.139	0.382	0.226	17.09	16.99	0.423	0.204	0.372	0.189
LDSR-S2 [18]	1	0.492	0.276	0.349	0.452	20.44	20.87	0.447	0.387	0.166	0.009
Real-ESRGAN [11]	1	0.458	0.300	0.440	0.456	20.34	20.67	0.444	0.388	0.168	0.016
SEN2SR Lite [17]	1	0.457	0.294	0.484	0.462	20.47	20.88	0.429	0.415	0.156	0.005
Bicubic	1	0.408	0.273	0.576	0.450	20.34	20.78	0.417	0.426	0.157	0.007
Bicubic (8-date median)	8	0.405	0.268	0.578	0.449	20.10	20.58	0.381	0.446	0.173	0.008

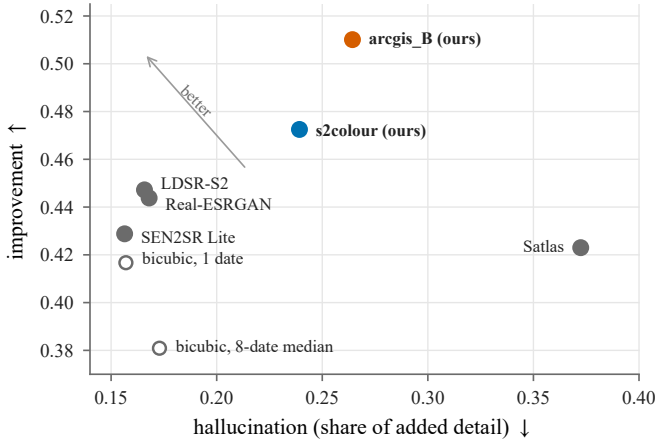


Fig. 2. opensr-test improvement (correct added detail) versus hallucination on the seven test sites.

canopy contains few sharp edges for any model to reconstruct. Snow-covered terrain is the most difficult case for all models (maximum edge-F1 0.393), since snow cover in the reference and in the input frequently differs. The hallucination score also varies with land cover. In the paddy landscape of Thanjavur, the score of *s2colour* is half that of *arcgis_B* (0.132 compared with 0.266), while both models reach their highest values (0.33–0.39) in the dense built-up chips of Hyderabad, Varanasi and Barmer.

C. Seasonal robustness and number of input dates

When the Sentinel-2 acquisitions are taken approximately six months from the reference capture date (Table IV), edge-F1 decreases by about 0.07 and LPIPS increases by about 0.03 for both models. Part of this decrease reflects actual changes in land cover between the input and the reference. Under these conditions both models still exceed the in-season scores of all published baselines. Table V reports results for the same weights with 1, 2, 4 and 8 input dates. With a single date, both models reach an edge-F1 of 0.60–0.62, higher than any single-image model (maximum 0.49), which indicates that most of the improvement results from training on Indian data. Additional dates yield a consistent, moderate gain (edge-F1 +0.02 to +0.04, LPIPS −0.03 to −0.04 from one to eight dates), most of which is obtained with four dates, in agreement with WorldStrat [7]. A single cloud-free acquisition

is therefore sufficient for rapid-response applications, at a measured cost of approximately 0.03 in LPIPS.

D. Constraint during training versus after training

s2colour retains most of the structural quality of *arcgis_B* (edge-F1 0.638 compared with 0.642, LPIPS 0.193 compared with 0.182), while its reflectance error with respect to the input decreases from 0.095 to 0.022, its spectral angle from 4.72° to 1.15° and its hallucination score from 0.264 to 0.239. Applying (1) to the output of *arcgis_B* after training produces a closer agreement with the input block means but a substantial loss of detail (Table VI, Fig. 3). Edge-F1 decreases to 0.528, closer to bicubic interpolation than to either proposed model, and LPIPS increases to 0.277. Because the colours of *arcgis_B* deviate considerably from Sentinel-2 (block error 0.036 in reflectance), the constraint applies large low-frequency corrections, which, superimposed on detail generated without them, appear as granular noise. When the constraint is present during training, the detail is retained without an increase in hallucination. The same difference is observed off-season (edge-F1 0.566 compared with 0.471, LPIPS 0.224 compared with 0.302).

E. Radiometric validation against $VEN\mu S$

Because the ArcGIS reference is a rendered basemap acquired on a different date, it cannot be used to validate band values. The SEN2VEN μS dataset [10], which pairs Sentinel-2 with same-day VEN μS surface reflectance at 5 m, is used instead. Its Kudaliar site in Telangana is located 215 km from the nearest training site and 75 km from the nearest test site. All 185 valid patches of 2019-08-27 (1.28×1.28 km each) were evaluated after averaging the 2.39 m output to the 5 m grid.

The band error of *s2colour* is 0.0081 (Table VII), compared with 0.0040 for the discrepancy between the two sensors and 0.058 for *arcgis_B*. The error of *s2colour* is thus within 0.004 of the inter-sensor discrepancy and approximately seven times lower than that of the unconstrained model. Bicubic interpolation obtains the best scores on all metrics in this comparison. This is expected, since VEN μS averaged to 10 m and upsampled again reproduces itself at 44.0 dB PSNR; the reference therefore contains little information beyond 10 m, and any added detail is counted as error (Section IV). The comparison consequently validates the band values but not

TABLE III
RESULTS PER SITE, IN-SEASON (400 TILES EACH): EDGE-F1 ($\uparrow$) AND COLOUR-NORMALISED LPIPS ($\downarrow$). BEST VALUE PER SITE IN BOLD.

Site	cover	edge-F1 $\uparrow$							LPIPS $\downarrow$						
		arcgis_B	s2colour	Satlas	LDSR	SEN2SR	R-ESRGAN	bicubic	arcgis_B	s2colour	Satlas	LDSR	SEN2SR	R-ESRGAN	bicubic
Hyderabad	urban	0.775	0.784	0.649	0.593	0.554	0.568	0.495	0.158	0.160	0.341	0.398	0.590	0.491	0.696
Varanasi	urban	0.714	0.692	0.610	0.451	0.418	0.440	0.377	0.145	0.163	0.374	0.359	0.503	0.434	0.591
Barmer	desert	0.739	0.717	0.584	0.527	0.488	0.482	0.442	0.164	0.176	0.344	0.353	0.505	0.470	0.605
Shillong	hills	0.708	0.710	0.581	0.524	0.487	0.492	0.433	0.177	0.174	0.372	0.395	0.519	0.492	0.621
Thanjavur	paddy	0.634	0.613	0.502	0.529	0.490	0.446	0.418	0.190	0.204	0.363	0.319	0.412	0.449	0.554
Wayanad	forest	0.535	0.554	0.413	0.536	0.499	0.504	0.451	0.172	0.188	0.416	0.254	0.388	0.316	0.443
Lachung	snow	0.389	0.393	0.327	0.285	0.263	0.276	0.241	0.266	0.286	0.463	0.365	0.472	0.427	0.521

TABLE IV
OFF-SEASON RESULTS: SENTINEL-2 ACQUISITIONS APPROXIMATELY SIX MONTHS FROM THE REFERENCE CAPTURE DATE (2,798 TILES; OPENSr-TEST ON CHIPS).

Model	edge-F1 $\uparrow$	grad $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	improv. $\uparrow$	omis. $\downarrow$	halluc. $\downarrow$
arcgis_B	0.570	0.288	0.214	0.400	0.455	0.268	0.277
s2colour	0.566	0.281	0.224	0.382	0.419	0.334	0.247
Satlas	0.486	0.107	0.406	0.199	0.405	0.246	0.348
Bicubic (8-date)	0.350	0.219	0.590	0.407	0.354	0.441	0.205

TABLE V
EFFECT OF THE NUMBER OF INPUT DATES (SAME WEIGHTS, IN-SEASON, 2,800 TILES).

Model	dates	edge-F1 $\uparrow$	grad-corr $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
arcgis_B	1	0.620	0.292	0.216	0.444
	2	0.632	0.323	0.182	0.449
	4	0.646	0.352	0.179	0.471
	8	0.642	0.356	0.182	0.473
s2colour	1	0.600	0.261	0.232	0.415
	2	0.619	0.302	0.198	0.418
	4	0.637	0.337	0.194	0.444
	8	0.638	0.344	0.194	0.447

the super-resolved detail. Fig. 4 shows that the structures reconstructed by s2colour are consistent in position with those in VEN μ S and Sentinel-2.

F. Transfer to a non-Indian dataset

On the opensr-test dataset of ten Spanish urban scenes, with a 2.5 m national orthophoto as reference and without any adaptation to Spain (Table VIII), s2colour obtains a higher improvement score than SEN2SR (0.255 compared with 0.193), a lower hallucination score (0.289 compared with 0.386) and a reflectance error approximately four times lower (0.0066 compared with 0.0289). Its spectral angle is slightly higher, which is attributable to the larger amount of added detail.

G. Geometric consistency

Misregistration between the stacked acquisitions would degrade the geometric accuracy of the output. On 39 test chips, each acquisition was co-registered to the eight-date median by phase correlation. The mean offset is 0.027 px (0.26 m), and the largest of 312 measurements is 0.201 px. The acquisitions are therefore sufficiently well aligned for a 2.39 m output, and the remaining sub-pixel offsets constitute the information exploited by multi-image SR.

TABLE VI
CONSTRAINT APPLIED DURING TRAINING (s2colour) VERSUS AFTER TRAINING (arcgis_B + CONSTRAINT); 2,800 TILES, 117 CHIPS, IN-SEASON. Δ : s2colour MINUS (arcgis_B + CONSTRAINT), WITH PAIRED BOOTSTRAP 95% CONFIDENCE INTERVAL (1,000 RESAMPLES). VALUES FROM A RE-RUN OF THE MODELS IN TABLE II, WHICH AGREE TO WITHIN 0.001. BOLD: THE BETTER OF THE TWO CONSTRAINED MODELS.

	s2colour	arcgis_B+lock	arcgis_B	Δ [95% CI]
edge-F1 $\uparrow$	0.637	0.528	0.642	+0.110 [0.107, 0.112]
LPIPS $\downarrow$	0.193	0.277	0.182	-0.084 [-0.086, -0.083]
grad-corr $\uparrow$	0.344	0.304	0.356	+0.040 [0.038, 0.041]
SSIM $\uparrow$	0.446	0.374	0.473	+0.072 [0.070, 0.074]
improv. $\uparrow$	0.472	0.415	0.510	+0.057 [0.050, 0.063]
omission $\downarrow$	0.288	0.347	0.226	-0.059 [-0.067, -0.051]
halluc. $\downarrow$	0.240	0.237	0.264	+0.002 [-0.002, 0.007]
block err. $\downarrow$	0.0104	0.0006	0.0359	+0.0099 [0.0096, 0.0101]



Fig. 3. High-structure test tiles from Barmer, Hyderabad, Lachung and Shillong (one per row). Columns: reference; s2colour; arcgis_B with the constraint applied after training; arcgis_B. Reference imagery © Esri and its providers.

The opensr-test spatial score reports offsets of 0.13–0.18 px for the proposed models. This score, however, interprets added high-frequency content as displacement: a control image composed of the bicubic upsampling and the true high-frequency component of the reference, which is correctly positioned

TABLE VII

COMPARISON WITH SAME-DAY VEN μ S AT 5 M, KUDALIAR, TELANGANA (185 PATCHES). BAND ERROR: MEAN ABSOLUTE REFLECTANCE DIFFERENCE OVER B04, B03 AND B02. SENSOR DISCREPANCY: SENTINEL-2 VERSUS VEN μ S AT 10 M.

	band err.↓	B04 / B03 / B02	angle (°)↓	PSNR↑	edge-F1↑	LPIPS↓
Bicubic	0.0043	0.0051 / 0.0042 / 0.0035	1.64	35.46	0.867	0.053
s2colour	0.0081	0.0086 / 0.0082 / 0.0076	2.48	29.98	0.834	0.218
arcgis_B	0.0581	0.0676 / 0.0653 / 0.0413	5.60	15.13	0.707	0.344
Sensor discrepancy	0.0040	—	1.53	—	—	—



Fig. 4. Kudaliar (VEN μ S comparison), central 640m of four patches. Columns: Sentinel-2 at 10m; bicubic at 5m; s2colour at 5m and 2.5m; arcgis_B at 2.5m; VEN μ S at 5m. VEN μ S data: CNES/Theia, CC BY-NC 4.0.

by construction, obtains 0.116. After a Gaussian blur with $\sigma = 1.5$ px, which removes fine detail but preserves translation, s2colour is offset by 0.010px from its input (control: 0.021). arcgis_B exhibits a systematic offset of approximately 0.08px (0.75 m), predominantly in the north–south direction, which originates from the reference imagery used in training (Fig. 5). s2colour is therefore the appropriate model where exact registration with other Sentinel-2 products is required.

H. Error decomposition and confidence estimation

Relative to bicubic interpolation, the proposed models reduce the omission score by 0.14–0.20; of this reduction, 0.06–0.09 corresponds to correct added detail and 0.08–0.11 to hallucinated detail (Table II). For practical use, the spatial location of the hallucinated component is relevant. Three candidate confidence layers were evaluated against the observed error on 840,000 pixels sampled from the 2,800 test tiles: the added-detail layer c , the edge strength of the output, and the per-pixel standard deviation between two predictions from disjoint halves of the input dates (Table IX).

For s2colour, the added-detail layer performs best, with an AUROC of 0.743 for identifying the 10% of pixels with the largest error. The 10% of pixels it ranks highest contain 33% of these errors, and the 10% it ranks lowest contain 3%. The AUROC remains 0.676 when pixels are stratified by edge strength, indicating that the layer captures more than the higher error at edges. For arcgis_B, none of the layers is a

TABLE VIII

OPENSr-TEST RESULTS ON SPANISH URBAN SCENES (10 SCENES, 2.5 M ORTHOPHOTO REFERENCE). NONE OF THE MODELS WAS ADAPTED TO SPAIN.

Model	reflect.↓	spectral (°)↓	spatial↓	improv.↑	omis.↓	halluc.↓
arcgis_B	0.0317	4.306	0.064	0.297	0.246	0.457
s2colour	0.0066	0.723	0.063	0.255	0.456	0.289
SEN2SR Lite	0.0289	0.671	0.006	0.193	0.421	0.386

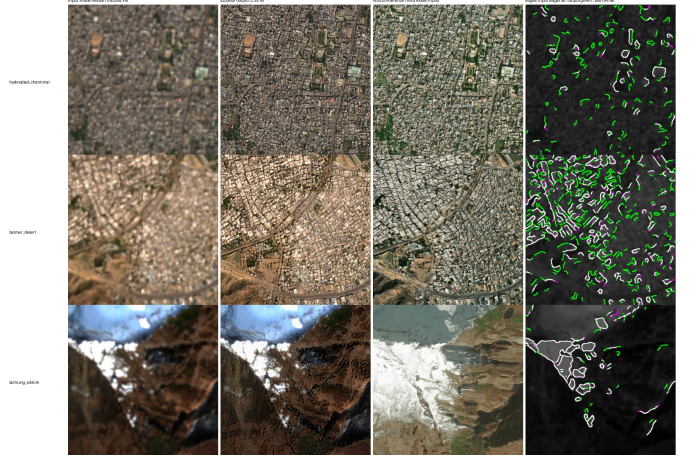


Fig. 5. Positional consistency of s2colour (Hyderabad, Barmer, Lachung). Columns: bicubic upsampling of the eight-date median input; s2colour; ArcGIS reference (not a model input); edges of input (magenta) and output (green), shown in white where they coincide. Reference imagery © Esri and its providers.

strong predictor, because its error is dominated by tile-specific colour and texture characteristics of the reference that cannot be inferred from the input.

VII. DISCUSSION

Applicability. The two models are intended for different uses of the same input data. arcgis_B provides the highest structural fidelity and is suited to visual interpretation and digitisation, particularly where recent VHR imagery is unavailable. s2colour provides comparable structural detail while preserving the Sentinel-2 band values at 10m: statistics computed over complete 10m cells equal those of the input (Proposition 1), the output is registered to the input within 0.010px, and a validated confidence layer indicates where detail is inferred. Since a single cloud-free acquisition yields most of the improvement, both models can be applied in rapid-response settings. Both models are available in a web-based GIS application (<https://viygan-2.pages.dev>) that provides confidence maps, spectral indices, zonal statistics and georeferenced export.

Effective resolution. The output is sampled at 2.39 m, but the input resolution is 10 m, and structure finer than a few metres is estimated from the sub-pixel information in eight acquisitions and from the learned prior. Hallucination is therefore reported explicitly, and a confidence layer is provided. For decisions at the scale of individual buildings, the output should be used to direct attention, and findings should be confirmed with VHR imagery.

TABLE IX

CONFIDENCE LAYERS EVALUATED AGAINST THE OBSERVED ERROR (840,000 PIXELS): AUROC FOR THE 10% LARGEST ERRORS, RANK CORRELATION WITH THE ERROR, AND AUROC WITHIN STRATA OF EQUAL EDGE STRENGTH. A VALUE OF 0.5 CORRESPONDS TO CHANCE.

Model	layer	AUROC $\uparrow$	rank corr. $\uparrow$	AUROC, equal edges $\uparrow$
arcgis_B	date-subset deviation	0.633	0.184	0.547
	edge strength	0.719	0.316	0.523
	detail added c	0.596	0.062	0.562
s2colour	date-subset deviation	0.647	0.211	0.564
	edge strength	0.723	0.324	0.525
	detail added c	0.743	0.326	0.676

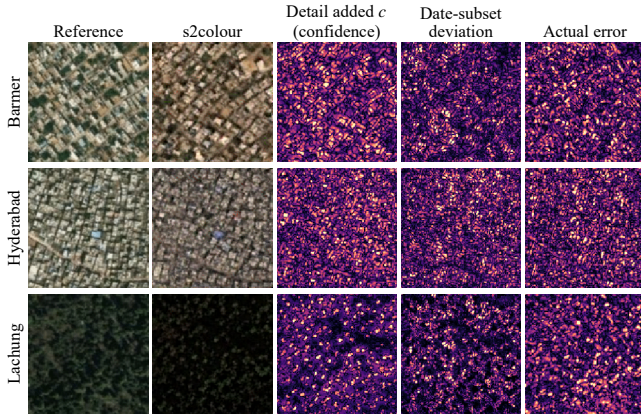


Fig. 6. Confidence layers for `s2colour` on three test tiles (Barmer, Hyderabad, Lachung). Higher values indicate more added detail, larger disagreement between date subsets, or larger error; each map is scaled to its 99th percentile. Reference imagery © Esri and its providers.

Reference data and constraint. The ArcGIS World Imagery reference is a rendered basemap acquired a median of 20 days from the input, so crop, water and shadow conditions can differ between input and reference. This affects all models equally, and band values were therefore validated separately against $\text{VEN}_{\mu\text{S}}$. The constraint ensures agreement with Sentinel-2 at the level of 10m cells within its residual (0.25–7/255 depending on the scene). It does not guarantee the correctness of individual 2.39m pixels, it follows the median of eight acquisitions rather than a single date, and it inherits the atmospheric-correction error of Sentinel-2 L2A. Because the input is stored on the ESA true-colour scale, reflectance above 0.3558 is saturated. Only the three visible bands are super-resolved, as the reference contains no other bands.

Scope. Although the test sites were not used in training, their land-cover types are represented in the training set; landscapes without a counterpart in the training data were not evaluated. Snow and closed forest are the most difficult cases. Each model was trained once on a single 8 GB GPU, with batch size and training length determined by the available hardware rather than by tuning. Following the Satlas authors, who found SR outputs ineffective as input to automatic classifiers [8], no claims are made regarding downstream task performance; its evaluation is left to future work.

VIII. CONCLUSION

This study adapted the Satlas eight-date ESRGAN to India using 103,936 Sentinel-2 L2A image pairs. On seven held-out sites spanning urban, agricultural, desert, forest, hill and snow terrain, the adapted model reduces LPIPS by more than half relative to the original model and outperforms all published models evaluated on structural and perceptual metrics. A parameter-free band-value constraint, combined with colour-invariant losses and a high-pass discriminator, allows the generator to retain nearly the same level of detail while every 10m cell preserves the Sentinel-2 B04, B03 and B02 values, as verified against same-day $\text{VEN}_{\mu\text{S}}$ reflectance. Applying the same constraint after training does not retain this detail. The PSNR penalty of the approach is bounded and measured, the hallucinated share of the added detail is quantified, and its spatial distribution is indicated by a validated confidence layer. The code is available at <https://github.com/Sarkar-Aveek/VIYGAN-2> (the `s2colour` code under PolyForm Noncommercial 1.0.0, the remainder under Apache 2.0), together with the weights of both models (CC BY-NC 4.0), all evaluation results and the scripts used to produce them.

APPENDIX A TRAINING DATA

Sources. The input consists of Copernicus Sentinel-2 L2A surface reflectance (bands B04, B03 and B02), obtained from the Copernicus Data Space Ecosystem and scaled according to the ESA true-colour product (reflectance/0.3558 $\rightarrow$ 1–255). L2A was preferred to L1C because it reduces the contrast difference to the reference from 20.5 to 12.5 (on a 0–255 scale). The reference consists of Esri ArcGIS World Imagery tiles at zoom level 17 (128 $\times$ 128 px, 2.39 m/px), resampled from 0.31–0.5 m acquisitions by WorldView-3 (32,787 training tiles), Legion (32,902), GeoEye-1 (8,967) and WorldView-2 (29,330). The reference imagery was used for non-commercial research only and is not redistributed.

Land-cover composition. The retained training tiles are distributed as follows: urban 43,419 (41.8%), farmland 20,263, forest 10,856, snow 8,590, desert 5,445, rocky terrain 4,788, hills 4,678, coast 2,897, river 2,050 and mountain 1,000.

Pair construction. For each site, the zoom-17 reference tiles closest to its centre are selected, excluding tiles dominated by water and tiles derived from imagery coarser than 1 m (for example, 84 Chennai tiles derived from 15 m imagery). Eight Sentinel-2 acquisitions within ± 120 days of the reference capture date are then selected in order of temporal proximity. Scenes with more than 30% cloud cover are rejected at catalogue level, and each remaining frame is compared with the per-pixel median of all candidate dates of the tile; a colour deviation above 20/255, or a structural deviation above $\max(5, 2.5 \times)$ the median of the tile, identifies the frame as affected by cloud or haze. Within each 16 $\times$ 16-tile block, tiles share the same dates where these are cloud-free, which avoids discontinuities in mosaics. Table X lists the subsequent filters. Of 132,879 reference tiles, 103,936 training pairs and 50 validation pairs were retained. The validation tiles originate

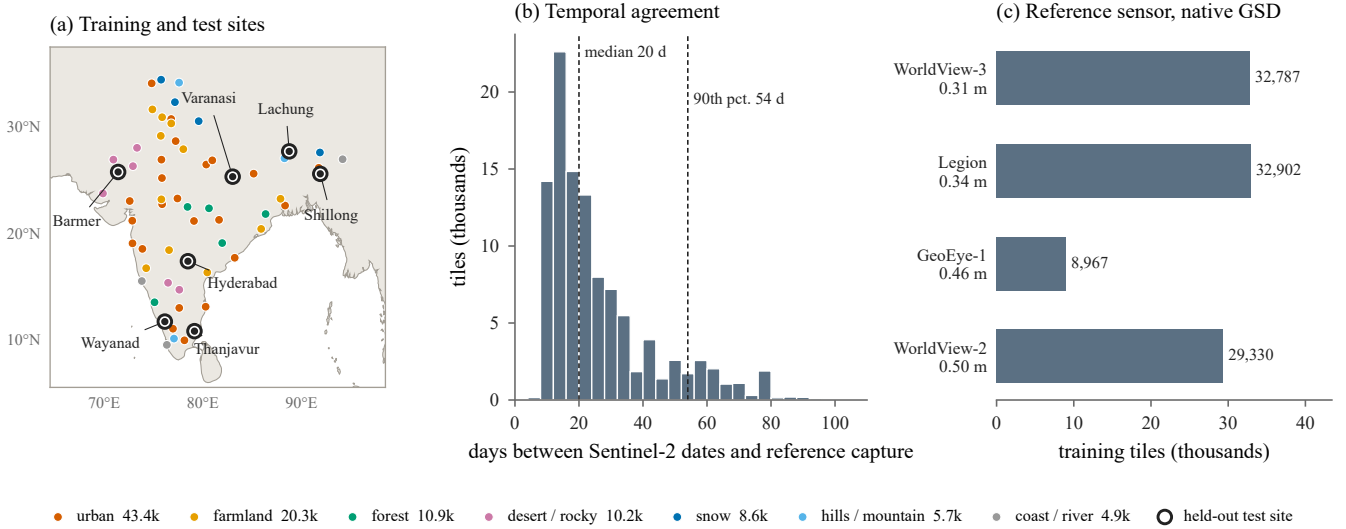


Fig. 7. Training and test data. (a) The 55 training sites, coloured by land cover (number of training tiles per class in the legend), and the seven held-out test sites (open circles); coastlines from Natural Earth. (b) Per tile, the median over its eight dates of the time difference between the Sentinel-2 acquisitions and the reference capture, for the training and in-season test tiles. (c) Sensor and native resolution of the imagery underlying the 2.39 m reference tiles.

TABLE X

FILTERS APPLIED DURING CONSTRUCTION OF THE TRAINING SET.

Filter	Rule	Removed
water	> 80% of the reference's 8×8 patches textureless	3,075
mismatch	reference (at 32 px) and Sentinel-2 median correlate < 0.2 ; land-cover change, or snow or cloud in one image only	6,411
site cap	sites used in the first Indian fine-tuning limited to 1,000 tiles each	18,674
no-data, duplicates	no-data pixels, incomplete stacks, repeated tiles	0

TABLE XI

TRAINING SEQUENCE AND SETTINGS. ALL RUNS: BATCH SIZE 12, ADAM, LEARNING RATES OF 5×10^{-5} (GENERATOR) AND 10^{-4} (DISCRIMINATOR), HALVED AT 80% WHERE INDICATED; ONE 8 GB LAPTOP GPU.

	India fine-tune	arcgis_B	s2colour
start data	Satlas G, new D 15 sites, L1C	India fine-tune G, D 55 sites, L2A	arcgis_B G, new high-pass D 55 sites, L2A
iterations	20,000	24,000 (halved at 80%)	24,000, then 17,322 from 18k
adversarial weight	0.1	0.05	0.05
edge loss	—	0.5	0.5
colour	reference	reference	Sentinel-2 (constrained)
wall time	2 h 42	2 h 47	3 h 05 + 2 h 02

from sites not seen by the initial model and were used only for monitoring and checkpoint selection.

Metadata audit. All 103,986 training and validation tiles have a recorded capture date, sensor, provider and resolution, and all originate from imagery with a resolution of 1 m or finer. The per-tile median time difference between the eight Sentinel-2 acquisitions and the reference capture has a median of 20 days over the training and in-season test tiles; 90% of tiles lie within 54 days (52 days for the training tiles alone), and the maximum is 106 days. Reference capture dates range from 2016 to 2026, and each reference is paired with Sentinel-2 acquisitions close to its own capture date.

APPENDIX B TRAINING SETTINGS

Table XI summarises the training sequence. Starting from the released Satlas eight-date weights, an initial fine-tuning was performed on 15 Indian sites with L1C input; *arcgis_B* was trained from this model on the complete L2A training set. The perceptual loss uses VGG19 layers conv1_2 to conv5_4, and the conditional discriminator receives 27 channels (the image and the eight upsampled dates). For *s2colour*, the

discriminator was pre-trained for 500 iterations with the generator fixed; training then ran for 24,000 iterations and was continued from the 18,000-iteration checkpoint for a further 17,322 iterations (two epochs). The EMA decay is 0.999.

The perception–distortion trade-off described in Section IV was also observed during training: between 6,000 and 24,000 iterations, the validation cPSNR of *arcgis_B* decreased from 22.91 to 22.79, while LPIPS (0.150 to 0.141), SSIM (0.500 to 0.518), edge-F1 (0.677 to 0.688) and gradient correlation (0.366 to 0.391) improved.

APPENDIX C ARCHITECTURE SELECTION

Prior to the main training, candidate architectures were compared on one 1.2×1.2 km scene at each of four Indian sites (Hyderabad, Varanasi, Thanjavur and Barmer, which subsequently became four of the seven test sites; the final checkpoints were selected on validation tiles only). Seven published models were evaluated as released, and eleven were fine-tuned for one epoch on the same 18,794 Indian tiles to provide an equal-budget comparison (Table XII). The multi-temporal ESRGAN obtained the lowest LPIPS at every site, and under equal training budget the Satlas ESRGAN variants

TABLE XII

ARCHITECTURE COMPARISON ON FOUR INDIAN SCENES (MEAN VALUES).
 “1 EP.”: FINE-TUNED FOR ONE EPOCH ON THE SAME 18,794 INDIAN
 TILES. INPUT: NUMBER OF DATES, OR THE EIGHT-DATE MEDIAN FOR
 SINGLE-IMAGE MODELS.

Model	input	LPIPS↓	cPSNR↑	SSIM↑
Satlas ESRGAN 8-frame, Indian fine-tune	8 dates	0.171	18.18	0.445
Satlas ESRGAN 2-frame, 1 ep.	2 dates	0.211	16.92	0.349
Satlas ESRGAN 8-frame, 1 ep.	8 dates	0.217	17.34	0.377
Satlas ESRGAN 1-frame, 1 ep.	1 date	0.218	16.99	0.308
Real-ESRGAN x4plus, 1 ep.	median	0.253	17.85	0.394
SwinIR-L [13], 1 ep.	median	0.280	18.01	0.387
LDSR-S2 [18], as released	1 date	0.415	17.90	0.326
Swin2-MoSE [14], 1 ep.	1 date	0.484	18.59	0.394
SEN2SR Lite [17], 1 ep.	1 date	0.495	17.98	0.340
SwinIR-L, as released	median	0.510	17.30	0.256
Bicubic	median	0.675	17.30	0.236

(0.211–0.218) outperformed Real-ESRGAN (0.253), SwinIR (0.280), Swin2-MoSE (0.484) and SEN2SR (0.495). Swin2-MoSE obtained the highest cPSNR (18.59), with an LPIPS of 0.484 and visibly smoothed output. The transformer models offered no advantage at this input size: SwinIR computes attention within 8×8 windows, of which a 32×32 px tile contains only 16, and single-image models cannot exploit multiple dates. The comparison is limited by the small number of scenes, by the use of L1C input for most models, and by the substantially longer training of the Indian ESRGAN compared with the one-epoch fine-tuning runs; the equal-budget results are reported for this reason.

APPENDIX D NUMBER OF CONSTRAINT ITERATIONS

With three iterations, the constraint leaves a mean block error of 0.0104 in reflectance (approximately 7/255) for *s2colour*, compared with 0.0006 when it is applied to *arcgis_B* after training, because the unconstrained output of *s2colour* is not trained to match the input brightness. The residual depends on the scene: 0.25/255 on validation tiles and approximately 3/255 on the VEN μ S patches. Additional iterations reduce the residual (on 60 patches, 3, 6 and 10 iterations leave 3.1, 0.16 and 0.01 on the 0–255 scale) but smooth the output (Fig. 8). Between 3 and 10 iterations, pixel values change by 3.7/255 on average and by up to 84/255, and the edge maps of the two versions agree with an edge-F1 of only 0.67. Each correction is distributed by bicubic interpolation, and repeated application progressively attenuates detail at the scale of one input cell. The released model uses three iterations.

ACKNOWLEDGMENT

This work contains modified Copernicus Sentinel data (2016–2026), accessed through the Copernicus Data Space Ecosystem. Reference imagery: Esri ArcGIS World Imagery (Vantor), used for non-commercial research. VEN μ S and Sentinel-2 patches were obtained from SEN2VEN μ S (CNES/Theia). The initial model is the Allen Institute for AI Satlas super-resolution model; training used BasicSR, and benchmarking used ESA opensr-test.

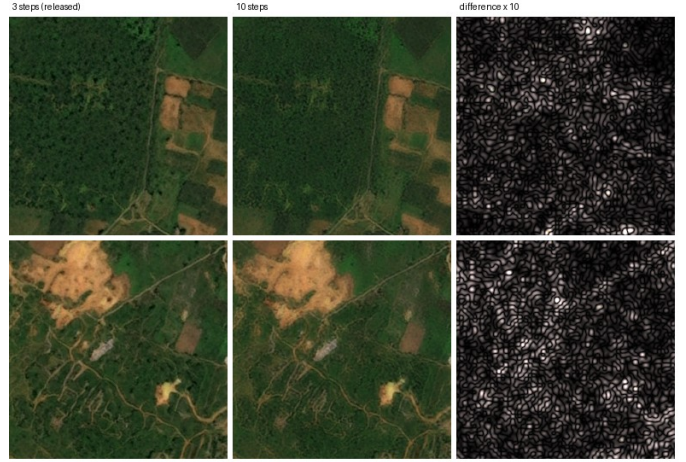


Fig. 8. Output of *s2colour* with 3 (released) and 10 constraint iterations, and their difference ($\times 10$), on two Kudaliar patches. Ten iterations match the input block means more closely but smooth texture and field boundaries.

REFERENCES

- [1] M. Drusch *et al.*, “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services,” *Remote Sens. Environ.*, vol. 120, pp. 25–36, 2012.
- [2] C. Ledig *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE CVPR*, 2017, pp. 4681–4690.
- [3] X. Wang *et al.*, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Proc. ECCV Workshops*, 2018.
- [4] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in *Proc. IEEE CVPR*, 2018, pp. 6228–6237.
- [5] M. Märtens, D. Izzo, A. Krzic, and D. Cox, “Super-resolution of PROBA-V images using convolutional neural networks,” *Astrodynamics*, vol. 3, pp. 387–402, 2019.
- [6] M. Deudon *et al.*, “HighRes-net: Recursive fusion for multi-frame super-resolution of satellite imagery,” arXiv:2002.06460, 2020.
- [7] J. Cornebise, I. Oršolić, and F. Kalaitzis, “Open high-resolution satellite imagery: The WorldStrat dataset – with application to super-resolution,” in *Proc. NeurIPS Datasets and Benchmarks*, 2022.
- [8] P. Wolters, F. Bastani, and A. Kembhavi, “Zooming out on zooming in: Advancing super-resolution for remote sensing,” arXiv:2311.18082, 2023.
- [9] C. Aybar, D. Montero, S. Donike, F. Kalaitzis, and L. Gómez-Chova, “A comprehensive benchmark for optical remote sensing image super-resolution,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, Art. no. 5003105, 2024.
- [10] J. Michel, J. Vinasco-Salinas, J. Inglada, and O. Hagolle, “SEN2VEN μ S, a dataset for the training of Sentinel-2 super-resolution algorithms,” *Data*, vol. 7, no. 7, Art. no. 96, 2022.
- [11] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in *Proc. IEEE/CVF ICCV Workshops*, 2021, pp. 1905–1914.
- [12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE CVPR*, 2018, pp. 586–595.
- [13] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin transformer,” in *Proc. IEEE/CVF ICCV Workshops*, 2021, pp. 1833–1844.
- [14] L. Rossi, V. Bernuzzi, T. Fontanini, M. Bertozzi, and A. Prati, “Swin2-MoSE: A new single image super-resolution model for remote sensing,” arXiv:2404.18924, 2024.
- [15] A. Retnanto *et al.*, “Beyond pretty pictures: Combined single- and multi-image super-resolution for Sentinel-2 images,” arXiv:2505.24799, 2025.
- [16] P. Kowaleczko *et al.*, “A real-world benchmark for Sentinel-2 multi-image super-resolution,” *Sci. Data*, vol. 10, Art. no. 644, 2023.
- [17] C. Aybar, J. Contreras, S. Donike, E. Portalés-Julià, G. Mateo-García, and L. Gómez-Chova, “A radiometrically and spatially consistent super-resolution framework for Sentinel-2,” *Remote Sens. Environ.*, vol. 334, Art. no. 115222, 2026.

- [18] S. Donike, C. Aybar, L. Gómez-Chova, and F. Kalaitzis, “Trustworthy super-resolution of multispectral Sentinel-2 imagery with latent diffusion,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6940–6952, 2025.
- [19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.