Better Data for Satellite Super Resolution

Miguel Castells, Jules Salzinger, Oliver Zendel

Austrian Institute of Technology Gmbh

Giefinggasse 4, 1210 Vienna, Austria

Reliable ML 2025
Workshop
San Diego
December 6th

Abstract:
Reliable satellite data is needed for many large-scale tasks in urban planning, agriculture, and disaster relief. However, high resolution satellite data is restricted or expensive. ESA's Sentinel-2 provides free satellite data with global coverage but only at a coarse level of detail. In this work we use super-resolution models trained to create high-resolution versions of Sentinel-2 data. We compare the feasibility of various CLIP embeddings to evaluate similarity between hallucinated satellite data and extend the existing S2-NAIP dataset. We automatically clean unreliable data and add new NIR band data. Our experiments show clear improvement in fidelity and quality of single image cross-sensor super resolution for satellite images.

Intro:
Earth Observation (EO) needs high-resolution satellite images
- Commercial sub-meter resolution have high costs
- Licensing, coverage limits, and security restrictions

ESA's Sentinel-2 [3] is a great source for data:
- Global coverage; 5-day revisit time; Open access
- **but** coarse GSD (one pixel relates to 10m x 10m)

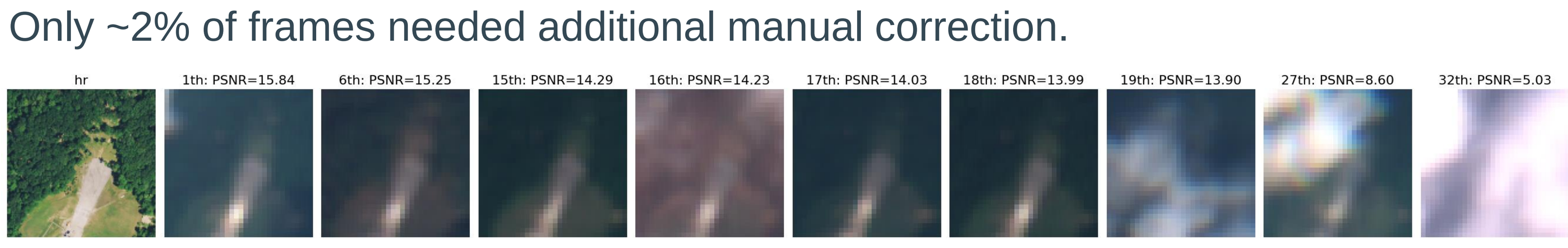
Super-resolution offers a way to train reliable ML models from lower-quality, upscaled satellite data but requires huge amounts of training data and is hard to evaluate.

Our contributions:
- Compare quality metrics based on different CLIP models.
- Refine S2-NAIP to improve cross-sensor LR/HR alignment for SISR.
- Add NIR bands to S2-NAIP.
- Train and evaluate efficient SISR using the improved S2-NAIP dataset

Improving S2-NAIP Dataset:
We extend the RGB S2-NAIP [1] dataset and add NIR band data. The original dataset includes 32 Sentinel2 snapshots for each NAIP[9] aerial image (various levels of cloud coverage, seasons, etc.). We devise a method to automatically detect the best low resolution (LR) Sentinel2 entry per high resolution NAIP entry:
Preprocessing:
- Remove NAIP tiles with black borders or missing time series (~1M pairs left).
- Remove Sentinel-2 frames with missing/corrupted data
- Keep the 16 best LR based on highest PSNR to drop cloudy / dark images

To reduce the domain gap (S2 vs. NAIP sensor/whitebalance etc.):
- Compute raw distance d(HR, LR) and distance d(HR_histm, LR) after histogram matching (scikit-image [10])
- Best combined score wins: Score_j = 0.7·d(HR, LR_j) + 0.3·d(HR_histmj, LR_j).

Only ~2% of frames needed additional manual correction.

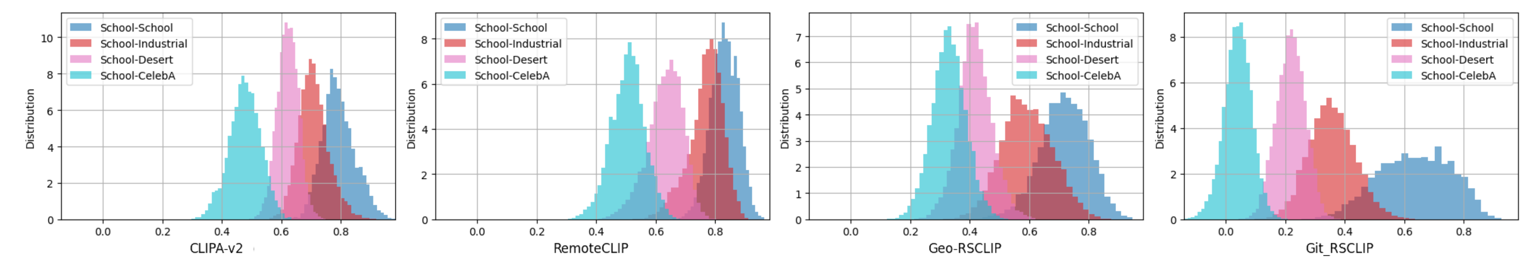


Example of automatic selection of best Sentinel2 LR sample relating to a HR NAIP image from the S2-NAIP dataset

Model	PSNR-RGBN (dB) ↑	SSIM-RGBN ↑	PSNR-RGB (dB) ↑	SSIM-RGB ↑	Git-RSCLIP-RGB ↑	LPIPS-RGB ↓
No Pretraining + **Finetuning**	25.885	0.548	27.860	0.594	0.670	0.358
Pretraining SEN2NAIPv2-HM	25.718	0.593	27.695	0.650	0.496	0.480
Pretraining SEN2NAIPv2-HM + **Finetuning**	**27.514**	0.588	**28.308**	0.603	0.734	0.343
Pretraining SEN2NAIPv2-UNet	27.112	**0.630**	27.912	**0.654**	0.492	0.470
Pretraining SEN2NAIPv2-UNet + **Finetuning**	27.432	0.587	28.249	0.601	**0.743**	**0.340**

Results evaluation metrics for the five trained single image super resolution ESRGAN[11] models

CLIP comparison:
PSNR and SSIM often fail to reflect perceptual quality when comparing super-resolution results. Idea: use cosine similarity of CLIP embeddings [12]. We evaluate the feasibility of four CLIP embedding versions for remote sensing applications: CLIPA-v2 [5,8], RemoteCLIP [7], Geo-RSCLIP [13], Git_RSCLIP [6]

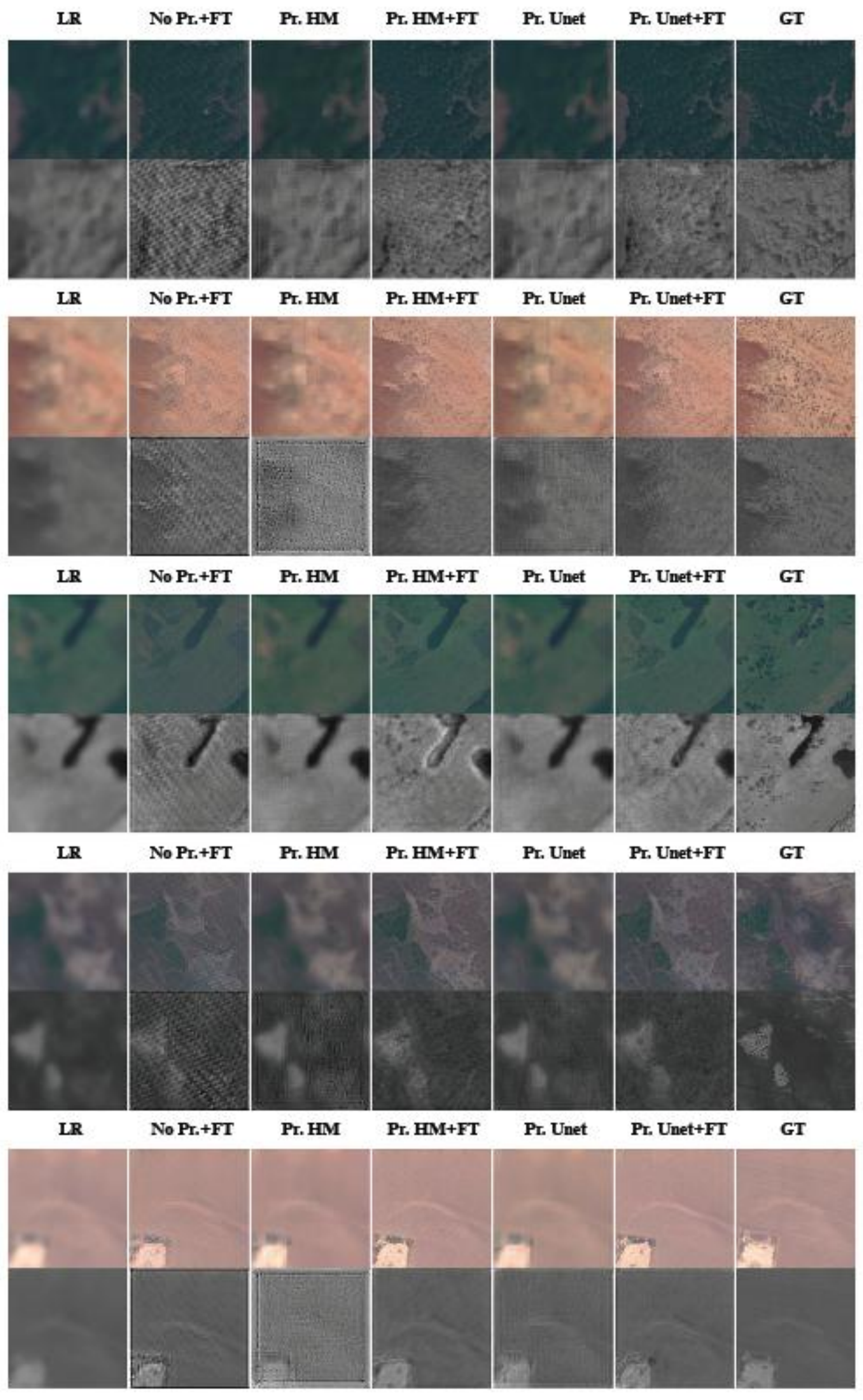


We expect good metrics to show high scores between remote sensing data of the same class (intra-class e.g. industrial area, schools, deserts) and large discrepancies towards out-of-distribution data (CelebA human faces). Also, semantically similar ground images should be closer than visually dissimilar scenes (e.g. school and industrial areas looking more similar than school and desert areas) **Git_RSCLIP** performed best while CLIPA-v2 showed weakest results.

Finetuning and influence of pretraining:
We trained new RGB-NIR SISR models based ESRGAN [11] using 2-step training [2,4]. Pretraining is done on SEN2NAIPv2 [1] (a single-sensor dataset) and finetune on our **improved S2-NAIP** cross-sensor dataset.

The comparison (see table below shows the impact of pretraining (SEN2NAIPv2-HM[1], SEN2NAIPv2-Unet[1]) and finetuning. Evaluations using the previously identified Git_RSCLIP corresponded best with visual comparisions:

Pretraining combined with finetuning yields the best results



Output images from all five ESRGAN[11] models

References

[1] Cesar Aybar et al. Sen2naip: A large-scale dataset for sentinel-2 image super-resolution. ScientificData, 11(1):1389, 2024.
[2] Simon Donike et al. Trustworthy super-resolution of multispectral sentinel-2 imagery with latent diffusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18:6940–6952, 2025.
[3] Matthias Drusch et al.Sentinel-2: Esa's optical high-resolution mission for gmes operational services. Remote sensing of Environment, 120:25–36, 2012.
[4] Liuyun Duan et al. Single Satellite Image Super-Resolution up to x8. In IGARSS2025, Brisbane (AU),Australia, August 2025.
[5] Samir Yitzhak Gadre et al. Datacomp: In search of the next generation of multimodal datasets. arXiv preprintarXiv:2304.14108, 2023.
[6] Chenyang Liu et al. Text2earth: Unlockingtext-driven remote sensing image generation with a global-scale dataset and a foundation model.IEEE Geoscience and Remote Sensing Magazine, pages 2–23, 2025.
[7] Fan Liu et al. Remoteclip: A vision language foundation model for remote sensing. IEEETransactions on Geoscience and Remote Sensing, 62:1–16, 2024.
[8] Clive Tinashe Marimo et al. Beyond the visible: Multispectral vision-language learning for earthobservation. arXiv preprint arXiv:2503.15969, 2025.
[9] U.S. Department of Agriculture, Farm Service Agency. National agriculture imagery program(naip) orthophotography. https://datagateway.nrcs.usda.gov, 2025. Accessed: 2025-08-27
[10] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne,Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: imageprocessing in python. PeerJ, 2:e453, 2014.
[11] Xintao Wang et al.. Esrgan: Enhanced super-resolution generative adversarial networks. In TheEuropean Conference on Computer Vision Workshops (ECCVW), September 2018.
[12] Piper Wolters et al. Zooming out on zooming in:Advancing super-resolution for remote sensing. 2023.
[13] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A largescale vision-language dataset and a large vision-language model for remote sensing. IEEETransactions on Geoscience and Remote Sensing, 62(1):1–23, 2024.6

Acknowledgments:
This research has been supported by the EMERALD project (Enhanced Multi-resolution Earth-observation using Robust and Advanced Learning for Environmental Dynamics), funded by the Austrian Research Promotion Agency (FFG).

Contacts:
miguel.castells@student-cs.fr
jules.salzinger@ait.ac.at
oliver.zendel@ait.ac.at

Project homepage:
github.com/ozendelait/better_s2_naip