MIMSS: A DATASET TO EVALUATE MULTI-IMAGE MULTI-SPECTRAL SUPER-RESOLUTION

Muhammed T. Razzak[*1], **Gonzalo Mateo-García**[*2], **Gurvan Lecuyer**[3]
Luis Gómez-Chova[2], **Yarin Gal**[1], **Freddie Kalaitzis**[1]
[1] OATML, Department of Computer Science, University of Oxford
[2] Image Processing Laboratory, University of Valencia
[3] European Space Agency

ABSTRACT

High resolution remote sensing imagery is used in a broad range of tasks, including detection and classification of objects. It is, however, expensive to obtain, while lower resolution imagery is often freely available and can be used for a range of social good applications. To that end, we curate a multi-image multi-spectral dataset for super-resolution of satellite images. We use PlanetScope imagery from the SpaceNet-7 challenge as the high resolution reference and multiple Sentinel-2 revisits of the same location as the low-resolution imagery. We provide baselines for both single image super-resolution and multi-image super-resolution. We also provide an ablation on how number of scenes, cloud cover and dynamism in different scenes in the dataset affect performance. Finally, we provide our code to obtain construct the dataset along with implementations of baselines for the community to build upon.[1]

1 INTRODUCTION

Generative Deep Learning has sparked a new wave of Super-Resolution (SR) algorithms that enhance the spatial resolution of images with impressive aesthetic results (Wang et al., 2020). Although the perceptual quality of those images is high, it is well-known that some of these SR models introduce artefacts into the SR image that are not present in real images (Bhadra et al., 2020). This limits the applicability of SR models to domains such as remote sensing where the safety and consistency are critical, e.g. for scientific instrumentation and decision making.

Super-resolution models can be divided in *single-image* super-resolution (SISR) and *multi-image* super-resolution (MISR) (aka *multi-frame* or *multi-temporal* super-resolution). The former uses as input only one low-resolution image while the later takes several low-resolution images from the same scene. MISR seeks to further constrain the ill-posed problem of SR by conditioning on several low-res input images (aka revisits). Therefore, it is expected of MISR to produce better SR images, to be more robust, and to produce fewer artefacts than SISR. In addition, MISR can be naturally applied in Earth observation since satellites often have frequent revisits of an area of interest. These multiple revisits can be fused with MISR to produce a super-resolved image. Despite its clear applicability, MISR has been scarcely applied in Remote Sensing and, as of yet, there are no studies that quantitatively compare MISR and SISR. Thus far, for remote sensing, MISR has been demonstrated only on RED and NIR bands of PROBA-V –a tiny fraction of the Sentinel-2 operation spectrum (Deudon et al., 2020).

In this paper, we introduce a dataset of multi-spectral multi-temporal satellite imagery from the European Space Agency's Copernicus Sentinel-2 (S2) archive, to test MISR. In particular, we train super-resolution models (both SISR and MISR) on the $10\,\mathrm{m}$ RGB bands of Sentinel-2 images, using as reference co-registered $4.77\,\mathrm{m}$ RGB PlanetScope images, acquired within the same two-month period. This setting differs from the vast majority of previous remote sensing applications of SR, where low-res images are obtained by artificially downsampling the high-res counterpart (Shermeyer & Van Etten, 2019). The main benefit of our setting is that the trained model can be applied on new

[1] Available at SpaceML.org

Figure 1: Different co-registered acquisitions from PlanetScope and Sentinel-2 from the SpaceNet 7 dataset. First row: PlanetScope RGB. Second and third rows: Sentinel-2 RGB revisits.

S2 RGB images to enhance their nominal resolution to $4.77\,\mathrm{m}$, i.e. it provides out-of-sample SR results without requiring simultaneous and co-registered VHR images. In addition, we provide baselines of not only MISR, but also SISR models on this dataset, in sec. 4.

The contributions of our work are summarized as follows:

1. We curate a new multi-temporal dataset from many revisits of Sentinel-2 imagery co-located with PlanetScope imagery, originally sourced from the SpaceNet-7 competition (Van Etten et al., 2021), which includes a geodiverse set of scenes from around the globe.
2. We present baseline results for both MISR and SISR on this dataset, along with some ablations of the dataset.
3. The code to construct similar datasets from Sentinel-2, along with the code to run the baselines on this dataset, is to be made publicly available.

The new multi-temporal dataset of Sentinel-2 and PlanetScope imagery, and the training, validation and test sets used in the experiments are presented in section 2. Section 3 describes the single-image and multi-image super-resolution methods analyzed in this work. Finally, the results of the baselines on the dataset in section 4.

2 DATASET

In order to learn a supervised super-resolution model to improve the spatial resolution of S2, we need higher-resolution images to be used as a reference. Since VHR (Very High Resolution) images (less than $10\,\mathrm{m}$) are not freely available, we restricted our search to pre-released publicly-available datasets of high-resolution images. Among those, we chose the recently launched Multi-temporal urban development SpaceNet dataset of PlanetScope images (also known as SpaceNet-7, see sec. 2.1) (Van Etten et al., 2021). Given this dataset, we acquired co-located time series of Sentinel-2 images for each PlanetScope acquisition (sec. 2.2). Subsection 2.3 has a brief analysis of the S2-Planet dataset as well as details about the different train-test splits that we used for the results.

2.1 PLANETSCOPE SPACENET-7 DATASET

SpaceNet-7 (Van Etten et al., 2021) has monthly time series of PlanetScope images over a two-year time span period for approximately 100 different areas of interest (AOI) all over the world (see Fig-

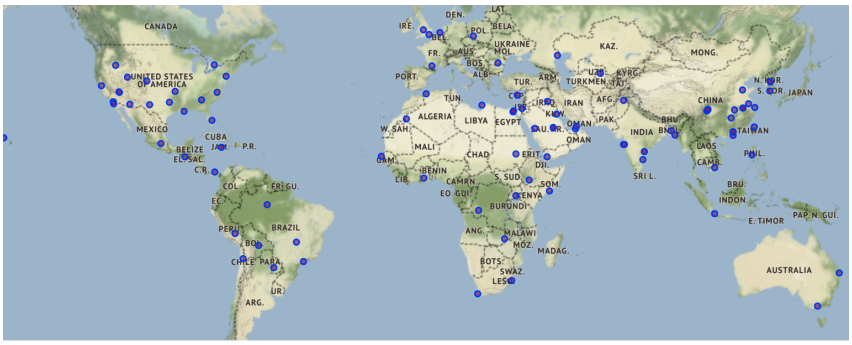


Figure 2: Location of SpaceNet-7 image time series. Figure taken from Van Etten et al. (2021).

ure 2). Images are provided at $4.77\,\mathrm{m}$ nominal[2] resolution with only three spectral channels (RGB) and size around 1000×1000 pixels. These images were sourced from Planet's global monthly basemaps which are mosaics of best scenes selected according to some quality metrics from PlanetScope Dove constellation as covered in Van Etten et al. (2021). Image values range from 0 to 255 and there is no information on the calibration or atmospheric correction of those images. In this study, we restricted to images from December 2019 and January 2020 from the training set (building footprints in the test set have not been released). In total there are 45 different PlanetScope scenes for each month.

2.2 SENTINEL-2 ACQUISITIONS

The Sentinel-2 mission consists of two twin satellites carrying the same multi-spectral optical sensor which acquires images on 13 different bands of the electromagnetic spectrum, from the visible to the short-wave infrared. The nominal spatial resolution of those images is different for each set of bands: 4 bands (visible and near infra-red) have $10\,\mathrm{m}$ resolution; 6 bands in the very near infrared and short-wave infrared have $20\,\mathrm{m}$ resolution, and the remaining 3 bands, which are used mainly for atmospheric correction, have a spatial resolution of $60\,\mathrm{m}$. Level 2A Sentinel-2 products consist of atmospherically corrected ortho-rectified 12-band images with bottom-of-atmosphere (BOA) calibrated reflectance. These images were downloaded from the European Space Agency (ESA) Open Access Hub. In order to obtain co-aligned time series of Sentinel-2 and PlanetScope images, we developed a custom pipeline which consists of the following steps:

1. Download all Sentinel-2 level 2A products overlapping with each of the 45 PlanetScope scenes over December 2019 and January 2020.
2. Crop all Sentinel-2 images to the PlanetScope scene bounds.
3. Reproject all bands of S2 to the coordinate reference system of PlanetScope products at $10\,\mathrm{m}$ spatial resolution.
4. When more than one S2 product was found for the same date and scene, we mosaiced those images.

2.3 DATASET ANALYSIS AND TRAINING SPLITS

The images come from 45 locations, with a geodiverse set of features including vegetation, bare soils (flats, hills, ridges), desert, urban, and agriculture infrastructure (see Appendix Table 3). The number of revisits between December 2019 and January 2020 range from 5 to 13 and the percentage of usable revisits ($< 50\%$ cloud coverage) ranges from 23% to 100%.

2.3.1 TRAINING, VALIDATION AND TEST SETS SPLITS

A well-thought split of the data into training and testing is critical to demonstrate the capacity of machine learning models to generalize. In remote sensing scenarios, extra-care must be taken to

[2]The resolution reported in the GeoTIFF metadata.

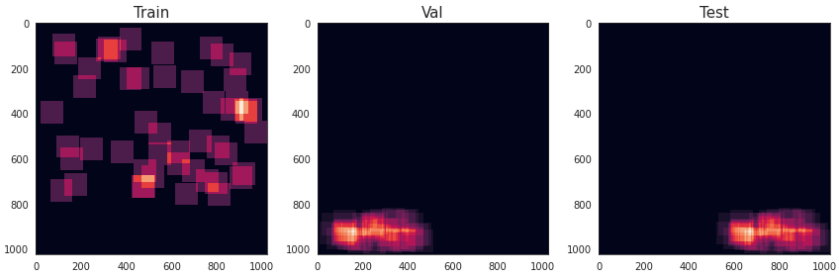


Figure 3: We utilised a within-scene split, which allocates the top 80% of a scene as a source of training patches; the bottom-left and right 10% for validation and testing patches respectively.

avoid train-test leakage due to spatial correlation. For instance, Ploton et al. (2020) recently showed that lack of consideration to spatial correlation lead to over-inflated results of ML models that monitored forest biomass. In this work, we test the models using two different training and testing splits. In the first approach, we split each scene in patches avoiding spatial overlap between patches in the different subsets; with this approach, we seek to explore the performance of the models in ideal conditions when training and testing patches come from similar distributions. Figure 3 shows the dataset partition for one scene following this approach. In the second approach, models are tested in images from the same location but different time periods (one month before). This split seeks to explore the capacity of the models to generalise to different time acquisitions. There are two scenes (0571E-1075N_2287_3888 and 0614E-0946N_2459_4406) that do not have PlanetScope images from one month before, so we excluded those images from the dataset.

3 BASELINES

In earth observation, there has been some recent work with Molini et al. (2020) and Deudon et al. (2020) tackling the MISR problem in Earth Observation, in single-band imagery. In particular, Deudon et al. (2020) was the first approach to tackle the different problems in MISR (input co-registration, fusion, and registration-at-the-loss) in an end-to-end manner, and with a small memory footprint (due to its reused fusion operator in the low-res domain). Since then, several deep learning approaches with refined architectures have repeatedly beaten the state-of-the-art in the Proba-V "post-mortem" leaderboard; most notably Salvetti et al. (2020).

This work modifies the HighRes-net architecture of Deudon et al. (2020) to the S-2 and PlanetScope RGB images described in sections 2.2 and 2.1, and provide this as the baseline MISR method. For a complete description of HighRes-net, we refer the reader to the original paper (Deudon et al., 2020).

For the Single-image super-resolution (SISR) baseline, we adapt the work of of Ledig et al. (2017), which provided a significant step forward in terms of photo-realistic and perceptually pleasing super-resolution. They introduced a far better CNN-based super-resolution network called Super-Resolution Residual Networks (SRResNet) and is a common baseline in traditional super-resolution tasks. In addition, when benchmarking MISR against SISR, it is important to define the Single-Image Selection strategy used in SISR. In particular, given the high Sentinel-2 revisit, one has to choose a revisit on which to do super-resolution using an SISR model. One could choose a random revisit, however, in order to have an stronger baseline, we choose the best revisit indicated by the lack of clouds.

3.1 SUPER RESOLUTION METRICS

For super-resolution, the primary quantitative metrics of performance are the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) (Wang et al., 2004) between the super-resolution image and the reference high-resolution image. Both measure the fidelity of the compared images, but with SSIM more focused on the structures contained in the images.

4 BASELINE RESULTS

In this section, we present the results of the Multi-Image Super-Resolution performance and image reconstruction quality. This section focuses on performance of the models in terms of Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (cf. sec. 3.1). These are the most commonly used metrics for measuring super-resolution performance and image reconstruction quality.

Table 1 shows the super-resolution performance of our evaluated models with the split described in section 2.3.1. *Bicubic* method indicates the scores if LR images are just upscaled using bicubic interpolation to match the HR image size. Bicubic interpolation and SRResNet (SISR) are computed using the best S2 revisit in terms of cloud coverage. Here we see that the performance of the MISR method (HighRes-net) is significantly higher than SISR (SRResNet) and Bicubic interpolation. Additionally, we see that scores calculated in training and testing splits do not differ much in the trained models (both in SSResNet and HighRes-net) which shows low overfitting.

Table 1: Average PSNR and SSIM scores (higher scores are better).

Subset	Bicubic		SRResNet		HighRes-net	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Train	17.85	0.612	27.90	0.827	**30.16**	**0.88**
Validation	18.11	0.70	26.06	0.78	**28.52**	**0.84**
Test	18.54	0.70	26.83	0.80	**29.40**	**0.85**

In addition, we tested the same model applied to acquisitions in a different time period but over the same regions. The results are shown in table 2. We notice a drop in performance across all methods, but the super-resolution models still exceed the performance of the bi-cubic upsampling method by a significant margin. Additionally, we still see that MISR outperforms SISR in all subsets.

Table 2: Average PSNR and SSIM scores for testing on a different time period.

Subset	Bicubic		SRResNet		HighRes-net	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Train	17.27	0.63	22.66	0.74	**23.28**	**0.76**
Validation	18.14	0.69	22.49	**0.73**	**22.56**	**0.73**
Test	16.99	0.66	22.12	0.73	**23.10**	**0.77**

4.1 STATIC VERSUS DYNAMIC SCENES ABLATION

One of the challenge when performing MISR on remote sensing data lies in the use of multiple revisits, which might show variability. In appendix C, we investigated the impact of the changing condition on the performance on HighRes-net model, and showed that using static scene allows achieving slightly better results in ideal conditions. However, using dynamic scenes helped to train a more robust model when dealing with data acquired at different time period.

5 CONCLUSION

In this work, we introduced a dataset for benchmarking super-resolution, in particular multi-spectral multi-image super-resolution, models on a real-world scientific application. We further provided results for baselines methods, using both MISR and SISR. The code to construct the dataset and run the baselines will be made publicly available upon de-anonymisation of the submission.

REFERENCES

Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *arXiv:2012.00646*, 2020.

Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E. Kahou, Julien Cornebise, and Yoshua Bengio. HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. *arXiv:2002.06460 [cs, eess, stat]*, February 2020. arXiv: 2002.06460.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017. doi: 10.1109/CVPR.2017.19.

A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli. DeepSUM: Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656, May 2020. ISSN 1558-0644. doi: 10.1109/TGRS.2019.2959248. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, September 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18321-y.

Francesco Salvetti, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge. Multi-image super resolution of remotely sensed images using residual feature attention deep neural networks. *arXiv preprint arXiv:2007.03107*, 2020.

Jacob Shermeyer and Adam Van Etten. The Effects of Super-Resolution on Object Detection Performance in Satellite Imagery. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1432–1441, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72812-506-0. doi: 10.1109/CVPRW.2019.00184.

Adam Van Etten, Daniel Hogan, Jesus Martinez-Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The Multi-Temporal Urban Development SpaceNet Dataset. *arXiv:2102.04420 [cs]*, February 2021. arXiv: 2102.04420.

Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861. Conference Name: IEEE Transactions on Image Processing.

ACKNOWLEDGEMENTS

We thank the Fontier Development Lab and the European Space Agency for funding this research. This research was also supported by an Nvidia grant and utilised Nvidia GPUs in this research.

A DATASET AREA OF INTEREST BREAKDOWN

Table 3: The subset (45) of SpaceNet-7 AOIs that we used in this work, acquired between December 2019 and January 2020. The high-level breakdown of the types of terrain contained in each scene shows the overall geodiversity of the dataset. Left to right: *% clouds* is the average cloud coverage (SCL=9) across all revisits; *desert, agri(culture), urban, veg(etation), bare (soils)* indicate the type of terrain; a *usable* revisit is at least %50 cloud-free.

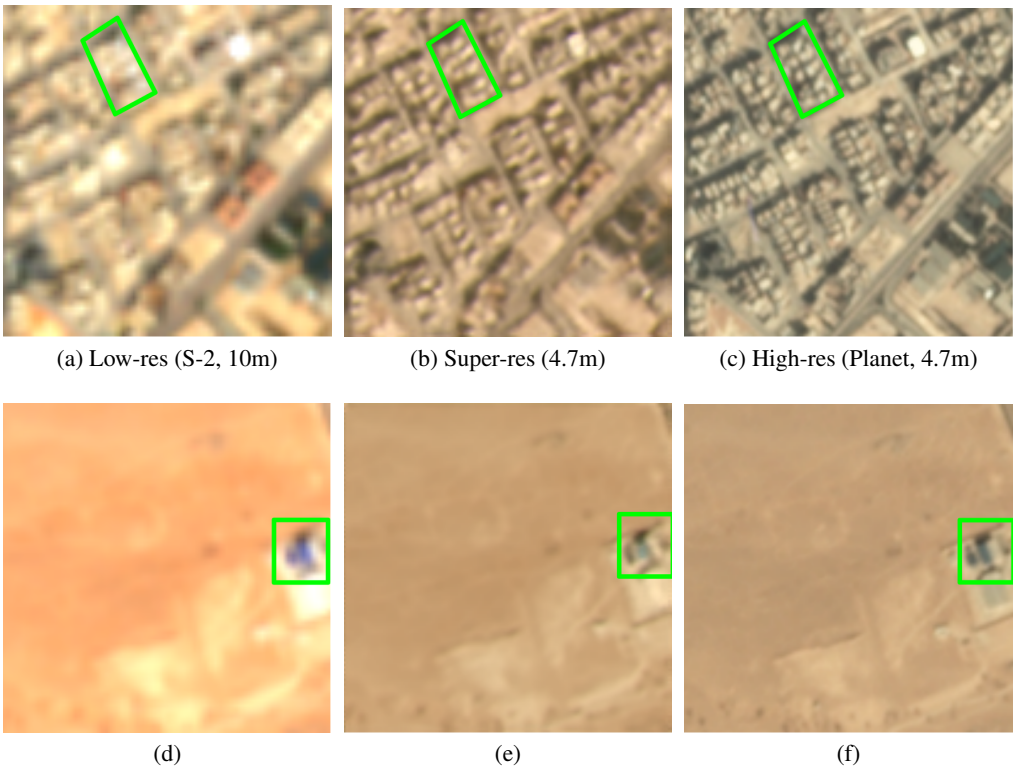
id	Scene	%clouds	desert	agri	urban	veg	bare	revisits		
								usable	total	%
1	0358E-1220N_1433_3310	70		1	1	1		8	13	62
2	1389E-1284N_5557_3054	69	1		1		1	8	13	62
3	0361E-1300N_1446_2989	66		1			1	8	13	62
4	1848E-0793N_7394_5018	66		1	1		1	7	13	54
5	0357E-1223N_1429_3296	65		1	1	1		6	13	46
6	1716E-1211N_6864_3345	62			1			5	13	38
7	1025E-1366N_4102_2726	55		1	1	1		5	13	38
8	1672E-1207N_6691_3363	53			1		1	5	13	38
9	1298E-1322N_5193_2903	56		1	1		1	4	13	31
10	1014E-1375N_4056_2688	49		1	1	1		4	13	31
11	1703E-1219N_6813_3313	58		1	1		1	3	13	23
12	1617E-1207N_6468_3360	24			1	1		3	13	23
13	1439E-1134N_5759_3655	61		1	1	1		5	10	50
14	0566E-1185N_2265_3451	96		1		1		6	7	86
15	0586E-1127N_2345_3680	83		1	1	1		6	7	86
16	1481E-1119N_5927_3715	81		1	1	1		6	7	86
17	0571E-1075N_2287_3888	81			1	1		6	7	86
18	1200E-0847N_4802_4803	80		1	1	1		6	7	86
19	1210E-1025N_4840_4088	95		1	1	1		5	7	71
20	1335E-1166N_5342_3524	84	1		1		1	5	7	71
21	1204E-1204N_4819_3372	74	1		1		1	5	7	71
22	0632E-0892N_2528_4620	67	1		1		1	5	7	71
23	1479E-1101N_5916_3785	67		1		1		5	7	71
24	0434E-1218N_1736_3318	84	1					4	7	57
25	1138E-1216N_4553_3325	71			1		1	4	7	57
26	0331E-1257N_1327_3160	68		1		1		4	7	57
27	1049E-1370N_4196_2710	59			1	1		4	7	57
28	1185E-0935N_4742_4450	33		1	1	1		2	7	29
29	0614E-0946N_2459_4406	29			1	1		2	7	29
30	1209E-1113N_4838_3737	100	1	1	1		1	6	6	100
31	0977E-1187N_3911_3441	100	1		1		1	6	6	100
32	1289E-1169N_5156_3514	99			1		1	6	6	100
33	0368E-1245N_1474_3210	67			1		1	6	6	100
34	1015E-1062N_4061_3941	100			1	1		5	6	83
35	1438E-1134N_5753_3655	92			1	1		5	6	83
36	1276E-1107N_5105_3761	91		1	1		1	5	6	83
37	1296E-1198N_5184_3399	87	1		1		1	5	6	83
38	0924E-1108N_3699_3757	67			1			4	6	67
39	0487E-1246N_1950_3207	98		1		1		3	6	50
40	1538E-1163N_6154_3539	63			1	1		3	6	50
41	1748E-1247N_6993_3202	57			1		1	3	6	50
42	1172E-1306N_4688_2967	56		1	1		1	3	6	50
43	1709E-1112N_6838_3742	44		1	1	1		3	6	50
44	0683E-1006N_2732_4164	58			1	1		2	5	40
45	0760E-0887N_3041_4643	33			1	1		2	5	40

B QUALITATIVE RESULTS OF MISR AND SISR ON THE DATASET

Fig. 4 shows a Sentinel-2 (low-res) image, a super resolved image using HighRes-net, and the PlanetScope (high-res) image. Particularly, the first row shows an urban area where the super-resolved image (middle) is significantly sharper than the low-res Sentinel-2 image (left); in this image, it is clear that counting buildings should be easier in the former than in the later.

C STATIC VERSUS DYNAMIC SCENES ABLATION

In this section, we present the result of a study conducted to measure the impact of the dataset on the performance of the MISR method. The super-resolution task relies on using multiple images acquired at different dates as input. Therefore, the revisits of the same scene can show significant changes caused by various factors such as vegetation, human activities, or weather events.



(a) Low-res (S-2, 10m) (b) Super-res (4.7m) (c) High-res (Planet, 4.7m)

(d) (e) (f)

Figure 4: Patch from a validation-set scene. How many buildings lie within the green polygon? Being more than just a pretty picture, the super-resolved output of HighRes-net 4b also better delineates the buildings in urban scenes, hence enabling downstream tasks like building segmentation, with improved accuracy compared to prediction on a single S2 image. This is evidenced qualitatively by the fact that the manual count of buildings in the green polygon in 4b is easier to perform than in 4a. Note that in both examples/rows, the spectra of the super-res output is similar to the high-res PlanetScope reference —an undesirable side-effect if the spectral information of the source low-res instrument is better than that of the high-res instrument.

We separated the dataset into two subsets for this experimentation: a static set containing scenes with revisits showing no or very few changes and a dynamic set containing scenes with revisits visible difference. For each scene, we computed the average pixel variance over the revisits. We used the median average pixel variance as a threshold to discriminate between static and dynamic scenes. The final static scene contained twenty-three scenes against twenty-two for the dynamic scene. Most of the changes observed in the dynamic scenes were due to weather effects such as cloud shadow, or changes in the illumination conditions.

We trained five MISR models on the static subset and five on the dynamic subset for this experimentation. Each model was then tested on both subsets individually, following the split described in section 2.3.1. We used the PSNR and SSIM metrics to evaluate the model performances. The results depicted in the following paragraph are the average value measured.

The Table 4 shows the results measured for both condition. $MISR_{stat}$ refers to the models trained using the static scenes, $MISR_{dyn}$ refers to the models trained using the dynamic scenes. On average $MISR_{stat}$ models reached a better PSNR on the static scenes, while the performances on the SSIM metrics are equal. When applied on the dynamic test set, both $MISR_{stat}$ and $MISR_{dyn}$ performances decrease and are almost equivalent. Overall, using static scenes for training allows achieving better results in this ideal condition.

As in section 4, we also tested the models applied on acquisitions from a different period but over the same regions. We used the test set without discriminating between static and dynamic scenes. The results are shown Table 5. We notice that the average performances dropped drastically across all models. However, the models trained using the dynamic scenes achieved slightly better perfor-

Table 4: Average PSNR and SSIM scores (higher scores are better).

Subset	$MISR_{stat}$		$MISR_{dyn}$	
	PSNR	SSIM	PSNR	SSIM
Static	28.62	0.85	28.44	0.85
Dynamic	28.00	0.86	28.01	0.85

mances than those trained on the static scenes. Models trained with dynamics scenes were exposed to more variability during the training step, which, we suspect, helps the models be better at generalization. We see that for both training sets, the performances decreased compared with models trained on the whole dataset (cf. tab. 2). It can be explained by the fewer data used for training, divided by two in both cases.

Table 5: Average PSNR and SSIM scores (higher scores are better).

Subset	$MISR_{stat}$		$MISR_{dyn}$	
	PSNR	SSIM	PSNR	SSIM
Test	19.17	0.74	19.76	0.76