MuS2: A Benchmark for Sentinel-2 Multi-Image Super-Resolution

Paweł Kowaleczko, Tomasz Tarasiewicz, Student Member, IEEE, Maciej Ziaja, Daniel Kostrzewa, Member, IEEE, Jakub Nalepa, Member, IEEE, Przemysław Rokita, Member, IEEE, and Michal Kawulok, Member, IEEE

Abstract—Insufficient spatial resolution of satellite imagery, including Sentinel-2 data, is a serious limitation in many practical use cases. To mitigate this problem, super-resolution reconstruction is receiving considerable attention from the remote sensing community. When it is performed from multiple images captured at subsequent revisits, it may benefit from information fusion, leading to enhanced reconstruction accuracy. One of the obstacles in multi-image super-resolution consists in the scarcity of real-life benchmark datasets—most of the research was performed for simulated data which do not fully reflect the operating conditions. In this letter, we introduce a new MuS2 benchmark for multi-image super-resolution reconstruction of Sentinel-2 images, with WorldView-2 imagery used as the high-resolution reference. Within MuS2, we publish the first end-to-end evaluation procedure for this problem which we expect to help the researchers in advancing the state of the art in multi-image super-resolution for Sentinel-2 imagery.

Index Terms—Super-resolution reconstruction, multi-image super-resolution, Sentinel-2, benchmark dataset.

I. INTRODUCTION

Super-resolution (SR) is aimed at reconstructing a high-resolution (HR) image from a single image or multiple low-resolution (LR) observations presenting the same area of interest. Multi-image SR (MISR) fuses multiple LR images, each of which contains a different portion of HR information. This allows for achieving higher reconstruction accuracy than relying on single-image SR (SISR) [1], but MISR is highly sensitive to the variability of the input images and their proper co-registration [2]. This poses a challenge when preparing the data for training and validation. Recent advances in satellite image SR include SISR [3] and MISR [4] techniques developed to overcome the problem of insufficient spatial resolution of Sentinel-2 (S-2) multispectral images (MSIs). They are composed of 13 bands, whose resolution ranges from 60 m ground sampling distance (GSD) to 10 m GSD [5].

A. Related Work

Most of the SR techniques have been evaluated relying on an artificial scenario—simulated LR images are obtained by downsampling and degrading an original image that is later treated as an HR reference. The similarity of the latter to the super-resolved outcome is then used to evaluate the SR performance. Unfortunately, such procedure does not reflect the real-life operating conditions, and methods that perform well for the simulated data are not necessarily effective when fed with original (i.e., not downsampled) images. In order to deploy SR in practice, it is crucial to properly validate the emerging techniques using real LR images coupled with an HR reference. Recently, several real-life SISR datasets have been elaborated which encompass pairs of original LR and HR images [6]. Some of the SISR techniques applied to S-2 were validated with PlanetScope [3] and WorldView-3 [7] images treated as a real HR reference, but these data have not been released for the community to serve as a benchmark. Preparing such datasets for MISR is much more costly and troublesome, as multiple LR images are necessary in this case. In 2019, European Space Agency organized an SR challenge and published a dataset with real scenes acquired by the Proba-V satellite, each of which contains an HR image (100 m GSD) coupled with at least nine LR images (300 m GSD). The dataset allowed for developing first MISR techniques underpinned with deep convolutional neural networks (CNNs), applied either to enhance the input LR images before their final multi-temporal fusion [8] or employed to learn the whole reconstruction process in an end-to-end manner [2]. The latter include HighRes-net that recursively combines latent LR representations to obtain the super-resolved image [9], as well as the residual attention MISR (RAMS) network equipped with the attention mechanism [10].

Very recently, the WorldStrat dataset was published which matches multiple S-2 images with SPOT images of 6 m GSD [11]. However, the problem of comparing the reconstruction outcome with the reference was not discussed there and it is not clear whether WorldStrat can be used as a benchmark for quantitative assessment of the SR outcome. As
demonstrated in this letter, this is not a straightforward task, especially when LR and HR images are captured by different satellites equipped with different sensors. In [7], SISR for S-2 images was assessed by comparing the outcome against WorldView-3 images. The authors observed that the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics do not correlate well with the quality of the reconstructed images—the highest scores were obtained for the techniques that produce blurred images in which the details were not reconstructed accurately. Contrary to that, the learned perceptual image patch similarity (LPIPS) [12] was reported to be suitable for evaluating SISR for remote sensing [13].

B. Contribution

In this letter, we address the problem of evaluating MISR for S-2 by introducing a new Multi-image Sentinel-2 SR (MuS2) benchmark. It is composed of a new dataset with WorldView-2 (WV-2) images used as an HR reference and the end-to-end validation procedure. Our contribution can be summarized in the following points.

- We publish a new MuS2 benchmark dataset[1] with 91 diverse scenes covering around 2500 km² (Fig. 1), each composed of at least 14 S-2 MSIs coupled with a single WV-2 MSI. MuS2 is supposed to serve as a test set for evaluating future advancements in S-2 MISR. Importantly, we publish the data preparation source code[2] to allow for reproducing the dataset and generating training data of the same characteristics (Section II).

- For each scene, we match all S-2 10 m bands with the corresponding WV-2 bands. Also, we report the results of 3× magnification obtained with the state-of-the-art MISR techniques compared with image interpolation algorithms treated as a baseline (Section III).

- We demonstrate that the LPIPS metric [12] is suitable for evaluating the MISR outcome, even if the reference HR images are acquired using a different satellite. To verify that, we have performed a mean opinion score (MOS) survey, whose results are discussed in the letter.

- We have introduced relevance masks which indicate the image regions that can be considered for evaluation relying on pixel-wise image similarity metrics.

II. THE MuS2 BENCHMARK

A. Data Source

Our benchmark dataset is composed of images acquired by S-2 and WV-2 satellites. The S-2 images were downloaded from the Copernicus Open Access Hub, and we accessed the WV-2 images published within the European Cities dataset[3].

The S-2 data are organized into tiles, each of which contains a 100 × 100 km MSI, and aggregated into products that are distributed at different processing levels. Here, we exploit Level-2A which includes the bottom of atmosphere reflectance

\[ N \times \text{Magnification factor} \alpha \]

The MuS2 dataset is available at [https://doi.org/10.7910/DVN/1JMRAT](https://doi.org/10.7910/DVN/1JMRAT). The data preparation source code is available at [https://codeocean.com/capsule/8131193/tree/v1](https://codeocean.com/capsule/8131193/tree/v1).

The European Cities dataset is freely available at [https://earth.esa.int/web/guest/-/worldview-2-european-cities-dataset](https://earth.esa.int/web/guest/-/worldview-2-european-cities-dataset).

Each WV-2 tile contains a panchromatic image of 0.4 m GSD and 8 spectral bands at 1.6 m GSD. These are C—coastal (400–450 nm), B—blue (450–510 nm), G—green (510–580 nm), Y—yellow (585–625 nm), R—red (630–690 nm), RE—red edge (705–745 nm), NIR1 (770–895 nm), and NIR2 (860–1040 nm). Based on the coverage between the spectral ranges of S-2 and WV-2 bands (measured with the Dice coefficient—\(D_C\)), we have selected four pairs: B02 with B (\(D_C = 0.8\)), B03 with G (\(D_C = 0.68\)), B04 with R (\(D_C = 0.68\)), and B08 with NIR1 (\(D_C = 0.92\)). For the remaining band pairs, the spectral coverage did not exceed \(D_C = 0.5\), so we do not consider them in our study.

B. Scene Selection and Data Alignment

For each WV-2 tile, we collected 14 or 15 S-2 images that entirely contain that tile or overlap with it. The images, both within the S-2 stacks, as well as with the WV-2 tiles, were co-registered at whole-pixel precision. We assemble the dataset using areas defined by seven different military grid reference system (MGRS) tiles. These areas represent diverse environments throughout Europe, such as mountains in Norway, plains in Germany and Belgium and a coastal region in Spain. The diversity of the chosen locations is illustrated in Fig. 1. Our data preparation procedure (see Fig. 2) operates on original S-2 and WV-2 images. The common area is determined based on the geographic coordinates retrieved from the metadata, and for each band, \(N\) LR images (\(I_{n}\)) are cropped and coupled with the cropped WV-2 image. The latter is subsequently downsampled to create the HR reference (\(I_{HR}\)) whose dimension is \(\alpha \times\) larger than those of \(I_{n}\).

In this way, we acquired 91 pairs of HR images, each of which is coupled with multiple LR images (every HR and LR image is composed of four bands). To verify whether the bands are correctly co-registered after cropping, we composed color images and inspected them against color artefacts (Fig. 3a, b). In order to check the co-registration correctness between LR images and the HR image, we have assembled checkerboard
mosaicked images and inspected them visually for all the scenes (Fig. 3). Finally, we combined multiple S-2 images for each scene (Fig. 3).

C. The Evaluation Procedure

The evaluated SR process is fed with $N$ input images for reconstructing a specific S-2 band to produce a single super-resolved image ($I_{SR}$) which is subsequently compared with the corresponding $I_{HR}$. We measure the similarity between $I_{SR}$ and $I_{HR}$ with the PSNR, SSIM, and LPIPS metrics that are commonly employed for assessing the SR quality [13]. $I_{in}$ and $I_{HR}$ are acquired by different sensors, hence the differences between $I_{SR}$ and $I_{HR}$ result not only from the reconstruction accuracy, but also from the characteristics of the imaging sensor and the temporal changes of the scene (due to different acquisition time) [14]. Also, as the $N$ input images are co-registered in the LR space at whole-pixel precision, the super-resolved $I_{SR}$ may be displaced up to $\alpha$ pixels in each dimension compared with $I_{HR}$. For Proba-V images (acquired with the same satellite), it was sufficient to co-register $I_{SR}$ to $I_{HR}$ and to compensate the brightness bias [15]. However, as the differences between S-2 and WV-2 images are more substantial than among Proba-V images, in addition to co-registering $I_{SR}$ to $I_{HR}$, we also match the histogram of $I_{SR}$ to $I_{HR}$ before the evaluation. Importantly, this does not convey any HR information from $I_{HR}$ to $I_{SR}$, but only compensates the low-frequency differences between them. Also, we do not modify $I_{HR}$, so that every $I_{SR}$ obtained using different techniques is always compared to the same reference image.

As proposed in [13], in addition to reporting the direct values of PSNR, SSIM, and LPIPS, we also present these scores in relation to the bicubic interpolation, and for all these three metrics we compute the balanced metric as:

$$B(I_{SR}) = \frac{1}{3} \left( \frac{S_p(I_{bic})}{S_p(I_{SR})} + \frac{S_s(I_{bic})}{S_s(I_{SR})} + \frac{S_l(I_{bic})}{S_l(I_{SR})} \right),$$

where $S_p(x) = \text{PSNR}(x, I_{HR})$, $S_s(x) = \text{SSIM}(x, I_{HR})$, $S_l(x) = \text{LPIPS}(x, I_{HR})$, and $I_{bic}$ is obtained by averaging all bicubically-upsampled $I_{SR}$’s in the scene. Hence, $B < 1$ means better performance compared with the bicubic interpolation, and $B > 1$ indicates the opposite case (for PSNR and SSIM the higher score indicates higher similarity, and for LPIPS the lower the score, the higher the similarity).

During our experiments (reported later in Section III), we observed that PSNR and SSIM are not particularly effective in assessing the reconstruction accuracy when $I_{in}$ and $I_{HR}$ originate from different satellites—the scores little depend on the employed SR or interpolation procedure. Even though the outcomes obtained using MISR techniques better reflect the image details, it can be seen from Fig. 3 that there are areas in which bicubic interpolation consistently prevails over all of the considered SR techniques (trained from real-life data) in terms of the mean squared error (MSE). This may result from the temporal differences between $I_{in}$’s and $I_{HR}$—in the regions where bicubic interpolation renders lower MSE than all the SR techniques, the HR reference can be regarded as inappropriate to assess the reconstruction quality. Based on that, for every scene we generate the relevance masks which indicate the regions in $I_{HR}$ that can be used for evaluating the SR outcome, and the metrics are computed only for the pixels that are not masked out.

III. EXPERIMENTS

The goal of our experiments was twofold: (i) to confirm that MuS2, including the evaluation procedure, is suitable for assessing the reconstruction accuracy, and (ii) to report the scores of the state-of-the-art SR techniques, which can be used as a baseline for future research.

In this study, we focus on evaluating MISR performed for 10 m S-2 bands matched with the HR reference obtained from WV-2. In order to verify that based on MuS2, we can assess how much the reconstructed details resemble those observed in $I_{HR}$, we consider three groups of methods: (i) MISR networks trained with real LR and HR Proba-V images, (ii) MISR networks trained using simulated LR data (obtained by degrading an original S-2 image, later treated as $I_{HR}$), and (iii) image interpolation techniques treated as a baseline. In [13], we demonstrated that RAMS and HighRes-net trained from Proba-V images are successful in super-resolving S-2 data, and they do not generate artefacts observed when these networks are trained using simulated LR images. To verify how such artefacts influence the scores for MuS2, we report the performance for the networks trained with Proba-V NIR, Proba-V Red, and simulated images.

Quantitative results obtained for four S-2 bands without and with the relevance masks are reported in Table [1]. It can be seen that without the relevance mask, the PSNR and SSIM scores little differ between each other within particular bands,
and they do not indicate the SR techniques to be better than interpolation. This is against what can be judged from visual inspection of the reconstruction outcome (Fig. 5). The images rendered by HighRes-net and RAMS trained over Proba-V images present more details than the interpolated images (which are quite blurry), and they are free from the artefacts present for the models trained from simulated data. It is worth noting that the reconstructed details are similar to those seen in the WV-2 image without any hallucination effect. However, these qualitative observations are quantitatively reflected only in the LPIPS values, while PSNR and SSIM are slightly worse for HighRes-net and RAMS—overall, both LPIPS and $B$ indicate that SR networks perform better than interpolation and they penalize for the artefacts. Apparently, the differences between $T_{\text{In}}$ and $T_{\text{HR}}$ images resulting from different image acquisition conditions prevail over the accuracy of reconstructing the details when local pixel-wise metrics like PSNR or SSIM are used, but they have smaller impact on the feature-based LPIPS metric. When the relevance masks are applied, all the metrics indicate the superiority of SR networks and their $B$ scores are aligned in a similar order as without the mask.

As shown in Table I, the reported metrics (without the
relevance masks) are not consistent in indicating the SR performance. In order to verify which of them is most reliable for assessing the reconstruction accuracy, we have prepared a MOS survey composed of 15 queries. We considered the cases in which PSNR, SSIM, and LPIPS metrics pointed to a different SR or interpolation outcome as the most (eight cases) or the least (seven cases) similar to $I_{HR}$. In each query, the participants of diverse background (including Earth observation professionals and individuals without any remote sensing experience) were presented such three images alongside $I_{HR}$, and asked to select the best or the worst image for retrieving some specific details (e.g., delineating the roads). In this way, we wanted to prevent the participants from being biased toward picking an image of the best perceptual quality, instead of the one whose details are most similar to those in $I_{HR}$. The results of the survey (averaged from over 160 responses) are reported in Table II—clearly, LPIPS correlates most with the human judgement both in picking the most or least accurate reconstruction outcome (retrieved with an SR network trained from Proba-V data and an interpolation technique, respectively). Overall, LPIPS can be used as a reliable indicator of reconstruction accuracy. The use of the balanced metric $\mathcal{B}$ introduces an additional stability, as the PSNR and SSIM scores are sensitive to the hallucination effect which potentially may not affect the LPIPS metric.

### Table II

| Query type          | Method | PSNR | SSIM | LPIPS | Total |
|---------------------|--------|------|------|-------|-------|
| SR networks (Proba-V) | 3.92   | 14.30| 65.67| 74.95 |       |
| SR networks (simulated) | 11.67 | 0.00 | 7.61 | 12.98 |       |
| Image interpolation | 1.81   | 4.97 | 0.00 | 6.78  |       |
| **Total** | 11.39 | 19.34| 64.18| **100.00** | |

| Query type          | Method | PSNR | SSIM | LPIPS | Total |
|---------------------|--------|------|------|-------|-------|
| SR networks (Proba-V) | 5.85    | 0.00 | 0.00 | 5.85  |       |
| SR networks (simulated) | 2.07  | 10.93| 6.71 | 19.71 |       |
| Image interpolation | 0.00   | 0.00 | 74.44| 74.44 |       |
| **Total** | 7.92  | 10.93| 81.15| **100.00** | |

### IV. Conclusions

In this letter, we introduced a new MuS2 benchmark for assessing the accuracy of MISR for S-2 images. It is composed of original S-2 data coupled with HR references obtained from WV-2. We have proposed the evaluation protocol based on a balanced score built upon PSNR, SSIM, and LPIPS metrics, and we demonstrated that it is suitable for assessing the reconstruction accuracy. Based on the MOS survey, we showed that LPIPS metric can be employed for assessing the similarity to the ground truth and that it is robust against variations resulting from acquiring images by different satellites. Also, we have elaborated the relevance masks which select the regions that can be used for pixel-wise evaluation performed with "traditional" PSNR and SSIM metrics.

Although we have shown that all of the S-2 10 m bands can be coupled with corresponding WV-2 bands, evaluating reconstruction of the remaining bands, potentially relying on panchromatic WV-2 images, remains an open issue. With our data preparation procedure, it is possible to crop all the bands which may help address this challenging problem in the future. Furthermore, here we consider the magnification factor of $3 \times$, but with the prepared procedure, other factors may be considered as well. Finally, even though the reconstruction outcomes obtained with the state-of-the-art techniques are of definitely higher quality than the interpolated images, they are still far from the HR reference. We expect that MuS2 used as a test set for assessing emerging SR techniques will guide the researchers toward developing more effective SR techniques.

### References

[1] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, “Image super-resolution: The techniques, applications, and future,” Signal Processing, vol. 128, pp. 389–408, 2016.

[2] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, “DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images,” arXiv preprint arXiv:1907.06490, 2019.

[3] M. Galar, R. Sesma, C. Ayala, L. Albitzis, and C. Aranda, “Learning super-resolution for Sentinel-2 images with real ground truth data from a reference satellite,” ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1, pp. 9–16, 2020.

[4] M. Razazk, G. Mateo-Garcia, L. Gómez-Chova, Y. Gal, and F. Kalaizit, “Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation,” arXiv preprint arXiv:2111.03231, 2021.

[5] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort et al., “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services,” Remote sensing of Environment, vol. 120, pp. 25–36, 2012.

[6] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, “Real-world single image super-resolution: A brief review,” Information Fusion, vol. 79, pp. 124–145, 2022.

[7] M. Beaulieu, S. Foucher, D. Haberman, and C. Stewart, “Deep image-to-image transfer applied to resolution enhancement of Sentinel-2 images,” in Proc. IEEE IGARSS. IEEE, 2018, pp. 2611–2614.

[8] M. Kaulukol, P. Benecki, S. Piekaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, “Deep learning for multiple-image super-resolution,” IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 6, pp. 1062–1066, 2020.

[9] M. Deudon, A. Kalaizit, I. Gaytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornelis, and Y. Bengio, “HighRes-net: Recursive fusion for multi-frame super-resolution of satellite imagery,” 2020.

[10] F. Salvetti, V. Mazzia, A. Khalil, and M. Chiaberge, “Multi-image super-resolution of remotely sensed images using residual attention deep neural networks,” Remote. Sens., vol. 12, no. 14, p. 2207, 2020.

[11] J. Cornelis, I. Orsolic, and F. Kalaizit, “Open high-resolution satellite imagery: The WorldStrat dataset—with application to super-resolution,” arXiv preprint arXiv:2207.06418, 2022.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in CVPR, 2018.

[13] P. Wang, B. Bayram, and E. Sertel, “A comprehensive review on deep learning based remote sensing image super-resolution methods,” Earth-Science Reviews, p. 104110, 2022.

[14] P. Benecki, M. Kaulukol, D. Kostrzewa, and L. Skonieczny, “Evaluating super-resolution reconstruction of satellite images,” Acta Astronautica, vol. 153, pp. 15–25, 2018.

[15] M. Märtens, D. Izzo, A. Krzic, and D. Cox, “Super-resolution of PROBA-V images using convolutional neural networks,” Astrodynamics, vol. 3, no. 4, pp. 387–402, 2019.

[16] M. Kaulukol, T. Tarasiwicz, J. Nalepa, D. Tyra, and D. Kostrzewa, “Deep Learning for Multiple-Image Super-Resolution of Sentinel-2 Data,” in Proc. IEEE IGARSS, 2021, pp. 3885–3888.

[17] C. Ledig, L. Theis, F. Huszár et al., “Photo-realistic single image super-resolution using a generative adversarial network.” in Proc. CVPR, vol. 2, no. 3, 2017, p. 4.
MuS2: A Benchmark for Sentinel-2 Multi-Image Super-Resolution (Mean opinion score survey)

Pawel Kowaleczko, Tomasz Tarasiewicz, Student Member, IEEE, Maciej Ziaja, Daniel Kostrzewa, Member, IEEE, Jakub Nalepa, Member, IEEE, Przemyslaw Rokita, Member, IEEE, and Michal Kawulok, Member, IEEE

TABLE I: Questionnaire sent to the respondents to determine the mean opinion score (MOS). The questions and images are in the same order as in the survey. There are 15 questions concerned with the reconstruction accuracy. For each image, we provide the metric used to determine the best (or worst) match alongside the method used to obtain that result (obviously, such information was not revealed to the respondents).

1. Which of the following images presents the urban area in the MOST DETAILED WAY?

   - Reference image
   - PSNR (HRN_SIM), 21 votes
   - SSIM (Lanczos), 9 votes
   - LPIPS (RAMS_NIR), 136 votes

2. Which of the following images is the MOST DIFFERENT from the reference image?

   - Reference image
   - LPIPS (Bicubic), 119 votes
   - SSIM (HRN_SIM), 29 votes
   - PSNR (HRN_NIR), 18 votes

3. Which of the following images presents the annotated woodless area in the MOST ACCURATE way?
   This area was annotated in red in the reference image (right).

   - Reference image
   - SSIM (Lanczos), 10 votes
   - LPIPS (HRN_NIR), 89 votes
   - PSNR (RAMS_SIM), 67 votes
TABLE I – Continued from the previous page

4. Which of the following images presents the SHAPE of annotated buildings in the WORST way?
The buildings of interest are annotated in red in the reference image.

Reference image
PSNR (HRN_SIM), 10 votes
LPIPS (NN), 139 votes
SSIM (RAMS_SIM), 17 votes

5. In which image the rural roads are presented in the CLEAREST (MOST DETAILED) way?
The roads of interest are annotated in red in the reference image.

Reference image
LPIPS (HRN_NIR), 69 votes
PSNR (Lanczos), 21 votes
SSIM (RAMS_NIR), 76 votes

6. In which image the buildings manifest THE LEAST LEVEL OF DETAIL?

Reference image
LPIPS (RAMS_SIM), 78 votes
SSIM (HRN_SIM), 50 votes
PSNR (HRN_NIR), 38 votes

7. In which of the following images the junction is presented in the MOST ACCURATE (DETAILED) way?

Reference image
PSNR (RAMS_SIM), 30 votes
SSIM (Lanczos), 34 votes
LPIPS (RAMS_NIR), 102 votes
TABLE I – Continued from the previous page

8. In which of the following images it is the MOST CHALLENGING to distinguish separate trees?

| Image | Method | Votes |
|-------|--------|-------|
| Reference image | SSIM (RAMS_SIM) | 12 votes |
| | LPIPS (Bicubic) | 150 votes |
| | PSNR (RAMS_NIR) | 4 votes |

9. In which of the following images the area of interest (annotated in red in the Reference image) is presented in the MOST DETAILED way?

| Image | Method | Votes |
|-------|--------|-------|
| Reference image | PSNR (Linear) | 3 votes |
| | LPIPS (RAMS_NIR) | 150 votes |
| | SSIM (Bicubic) | 13 votes |

10. Which of the following images presents the coastal area in the LEAST ACCURATE (LEAST DETAILED) way? The area of interest is rendered in red in the reference image.

| Image | Method | Votes |
|-------|--------|-------|
| Reference image | LPIPS (Lanczos) | 154 votes |
| | SSIM (RAMS_SIM) | 2 votes |
| | PSNR (HRN_SIM) | 10 votes |

11. Which of the following images presents the roads in the MOST DETAILED way? The roads of interest are annotated in red in the reference image.

| Image | Method | Votes |
|-------|--------|-------|
| Reference image | SSIM (HRN_RED) | 10 votes |
| | LPIPS (RAMS_NIR) | 126 votes |
| | PSNR (RAMS_RED) | 30 votes |
12. In which of the following images the urban area is presented in the LEAST DETAILED way?

![Reference image](image1)

- SSIM (RAMS_SIM), 11 votes
- LPIPS (NN), 151 votes
- PSNR (HRN_SIM), 4 votes

13. Which of the following images allows one to count the buildings in THE MOST CONVENIENT way?

The buildings of interest are annotated in red in the reference image.

![Reference image](image2)

- PSNR (RAMS_NIR), 22 votes
- SSIM (HRN_NIR), 43 votes
- LPIPS (HRN_SIM), 101 votes

14. In which of the following images the parking is presented in the LEAST DETAILED way?

The parking of interest is annotated in red in the reference image.

![Reference image](image3)

- LPIPS (Bicubic), 152 votes
- PSNR (RAMS_NIR), 8 votes
- SSIM (RAMS_SIM), 6 votes

15. Which of the following images presents the roads of interest MOST FAITHFULLY?

The roads of interest are annotated in red in the reference image.

![Reference image](image4)

- SSIM (HRN_RED), 63 votes
- LPIPS (RAMS_NIR), 66 votes
- PSNR (RAMS_SIM), 37 votes