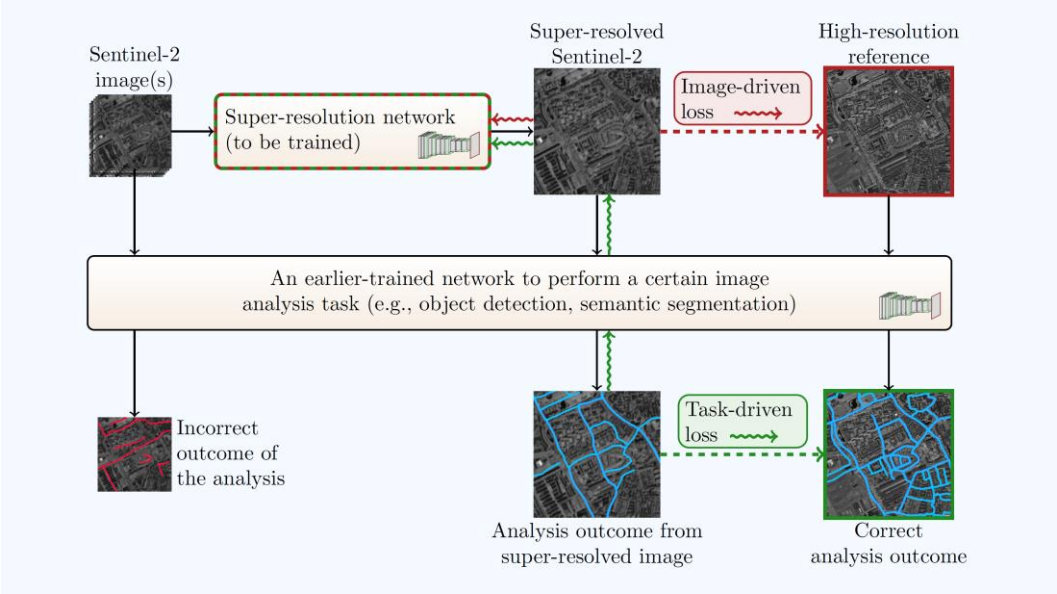
KP LABS
Mission Complete

esa

→ DISCOVERY



Task-driven super-resolution of Sentinel-2 images

Executive summary

OSIP Idea

Open Channel

Affiliation(s): KP Labs

Activity summary:
The goal of this project was to develop new techniques for super-resolving Sentinel-2 images underpinned with loss functions that embrace specific downstream tasks. In particular, road and building segmentation were employed to guide the training of deep networks for single-image and multi-image super-resolution. The results clearly indicate the merits of the developed approach in the context of real-world Sentinel-2 super-resolution.

Publishing Date: 14th May 2024
Contract Number: 4000141284/23/NL/GLC/ov
Implemented as ESA Express Procurement (EXPRO)

→ THE EUROPEAN SPACE AGENCY

ESA Discovery & Preparation
From breakthrough ideas to mission feasibility. Discovery & Preparation is laying the groundwork for the future of space in Europe
Learn more on www.esa.int/discovery
Deliverables published on https://nebula.esa.int

Super-resolution (SR) is aimed at generating a high-resolution (HR) image from low-resolution (LR) observation, being either a single image or multiple images of the same area of interest [11]. Recently, a lot of efforts have been invested into making the SR techniques suitable for enhancing real-world images that have not been degraded beforehand [2]. One of the ways to achieve that goal is to exploit real-world data for training that encompass original LR and HR images of the same area that were acquired independently. This is in contrast to employing simulated data for training and validation, in which LR images are obtained by downsampling and degrading the original images, later treated as an HR reference. However, using real-world data for training brings considerable challenges resulting from temporal changes between subsequent acquisitions and different characteristics of the imaging sensors [5]. State-of-the-art SR techniques are underpinned with deep networks that are commonly trained with loss functions that maximize the similarity between the reconstructed and reference images. However, pixel-wise image similarity is not sufficiently robust in such settings, leading to worse performance for real-world data than for the simulated ones. In this project, we addressed this problem by treating SR as a preprocessing step before performing specific image analysis tasks. The selected tasks were exploited as loss functions while training single-image SR (SISR) and multi-image (MISR) techniques, resulting in task-driven training. The activity was supposed to provide answers to the following questions:

1. How much can the analysis outcome be improved by super-resolving the input images with the models trained with a task-specific loss, compared with those trained using an image similarity loss?
2. Is the task-driven loss more robust against the variations resulting from different image acquisition conditions than the standard pixel-wise loss functions?
3. Considering the case when the simulated data are used for training models that are later applied for enhancing real-world images: is it better to use the task-driven loss in such cases than the pixel-wise loss functions?
4. What is the performance gap between MISR and SISR techniques trained in the task-driven manner?

The project was executed according to the initial plan presented in the proposal and the main achievements are summarized below.

- We have investigated several image analysis tasks, including road network segmentation [6], building network segmentation [7], keypoint detection (using Key.Net) [1], generic unsupervised image segmentation (Segment Anything) [4]. At first, these tasks were applied to images at different scales, including interpolated and super-resolved images. This led us to elaborating a methodology for assessing the suitability of such image analysis tasks for training the SR networks, which we will present at IGARSS 2024 [12].
- We have extended our MuS2 dataset with Sentinel-2 images coupled with WorldView-2 images with ground-truth masks of roads and buildings acquired from OpenStreetMaps [10]. We found using these masks during training to be more effective than relying on the masks obtained by segmenting the HR reference images. Furthermore, this opens a possibility of training real-world SR networks without HR reference images.
- We have developed the methodology to better understand the loss functions by inspecting their sensitiveness to the registration errors, as well as to the synchronized shifts of the input and target. Furthermore, we made attempts to investigate the optimization landscape by optimizing the input that is compared with a given target image.
- We found out that using a task-based loss function makes the SR network training heavily ill-posed—commonly, an almost perfect segmentation outcome can be retrieved from a noisy image which does not present any meaningful structures. Therefore, the task-based loss functions must be either employed for fine-tuning (after training the network with conventional image-based loss functions) or they must be combined with an image-based loss. Moreover, we employed task balancing for multi-task learning [9] which helped us

eliminate the problem of falling into a minimum defined by an “easier” loss function, without optimizing the remaining loss functions (or even at a cost of increasing their values). We have confirmed that by observing the loss functions, each of which is minimized during training.

- We have selected building segmentation and road segmentation tasks for training the SISR and MISR networks. For SISR, we selected the hybrid attention transformer (HAT) [3], and for MISR, we used residual attention multi-image super-resolution network (RAMS) [8]. After running many tests based on simulated and real-world images, we designed a set of around 30 final tests to answer the aforementioned research questions.

The obtained quantitative and qualitative results, reported in the D2 document, confirmed that task-driven training substantially improves the quality of the super-resolved images in terms of their value for image analysis tasks. As presented in Figures 1 and 2, road and building segmentation is much more effective from the images super-resolved using models trained in a task-driven way. This can be observed both for the simulated and original Sentinel-2 images. Also, the details are much more clear after running the task-driven training.

Overall, the performed activity allowed us to answer the aforementioned research questions (for details, see the D2.2 document):

1. *How much the analysis outcome can be improved by super-resolving the input images with the models trained with a task-specific loss, compared with those trained using an image similarity loss?*

   Response: From the obtained results, it is clear that introducing a task-based component into the loss function substantially improves the outcome of the task. This is much more evident for the real-world data, where the task outcome is extremely poor without applying the task-driven training. For the simulated data, the task outcome after regular training is slightly better, but introducing the task-driven loss increases the segmentation scores substantially.

2. *Is the task-driven loss more robust against the variations resulting from different image acquisition conditions than the standard pixel-wise loss functions?*

   Response: The question concerning the robustness was specifically aimed at real-world data in which the LR and HR images have been captured independently. The reference-based metrics (i.e., the similarity to the reference image) were fairly similar across different trainings, but taking into account the increase in image segmentation quality after task-driven trainings, it can be concluded that the task-driven training deals well with the differences between LR and HR images that result from the imaging conditions. When the SR outcomes generated after task-driven training (see Figures 1 and 2) are closely inspected, then it can be noticed that the enhanced details are achieved at a cost of some grid-like artifacts. Overall, the answer to this question is positive—task-driven training increases the robustness against the variations resulting from different image acquisition conditions.

3. *Considering the case when the simulated data are used for training models that are later applied for enhancing real-world images: is it better to use the task-driven loss in such cases than the pixel-wise loss functions?*

   Response: Comparing the scores rendered by the models trained using simulated data with regular and task-driven loss functions, it becomes clear that the task-driven training leads to similar reference-based scores, but the image segmentation scores are definitely

higher. However, in both cases some artifacts are visible and they are more severe than when the training is performed from the real-world data.

4. *What is the performance gap between MISR and SISR techniques trained in the task-driven manner?*

   Response: This question has been answered only partially, as it eventually occurred that the SISR model cannot be used to super-resolving the same scenes as the MISR model. The use of smaller patches in SISR caused the image similarity scores to be inflated, and eventually higher than for MISR. However, the visual quality of SISR is below that of MISR and also the segmentation scores were worse for SISR. Still, it is clear that the trends observed for MISR also hold for SISR and task-driven training is extremely helpful here.

In addition to that, we have demonstrated that a model trained from the simulated data (relying on a regular loss) can be fine-tuned using real-world data without the HR references. In such a case, only the segmentation maps may be used to compute the task-based loss which is combined with the consistency loss that prevents the network from falling into minima.

Overall, we have found the obtained results to be positive and in our opinion the problem of task-driven training is worth further investigation. The results show that the SR space is quite broad and using a single task (including the image-based similarity) may not be sufficient to evaluate SR techniques and guide their training effectively. In particular, we identify the following future research pathways:

1. Involving more tasks to stabilize the result and narrow down the SR space.
2. Enhancing the multi-task learning via exploiting better loss balancing procedures.
5. Combining the use of simulated data for training, coupled with the consistency loss that will allow for self-supervised training of SR networks, suitable for real-world applications.
6. The aforementioned actions should be aimed at reducing the artifacts while preserving the high performance of image analysis tasks.
7. Developing a robust SR evaluation protocol underpinned with image-based and task-based features. For that purpose, we plan to enhance our MuS2 dataset.

The initial results obtained in this project were described in a research paper that was accepted at IEEE IGARSS 2024 [12]. Also, they will be presented at SUREDOS conference organized by ESA in May 2024. In addition to that, we are planning to prepare a journal paper summarizing the obtained results.

References

[1] A. Barroso-Laguna and K. Mikolajczyk, “Key.Net: Keypoint detection by handcrafted and learned CNN filters revisited,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 698–711, 2022.

[2] 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.

[3] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, “Activating more pixels in image super-resolution transformer,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 22 367–22 377.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., “Segment anything,” arXiv preprint arXiv:2304.02643, 2023.

[5] P. Kowaleczko, T. Tarasiewicz, M. Ziaja, D. Kostrzewa, J. Nalepa, P. Rokita, and M. Kawulok, “A real-world benchmark for Sentinel-2 multi-image super-resolution,” Scientific Data, vol. 10, no. 1, p. 644, 2023.

[6] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, “Coanet: Connectivity attention network for road extraction from satellite imagery,” IEEE Transactions on Image Processing, vol. 30, pp. 8540–8552, 2021.
[7] V. Mnih, Machine learning for aerial image labeling. University of Toronto (Canada), 2013.
[8] F. Salvetti, V. Mazzia, A. Khaliq, 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.
[9] S. Vandenhende, S. Georgoulis,W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, “Multi-task learning for dense prediction tasks: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3614–3633, 2021.
[10] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, “OpenStreetMap: Challenges and opportunities in machine learning and remote sensing,” IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 1, pp. 184–199, 2020.
[11] 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.
[12] M. Ziaja, P. Kowaleczko, D. Kostrzewa, N. Longepe, and M. Kawulok, “Toward task-driven satellite image super-resolution,” in Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2024.

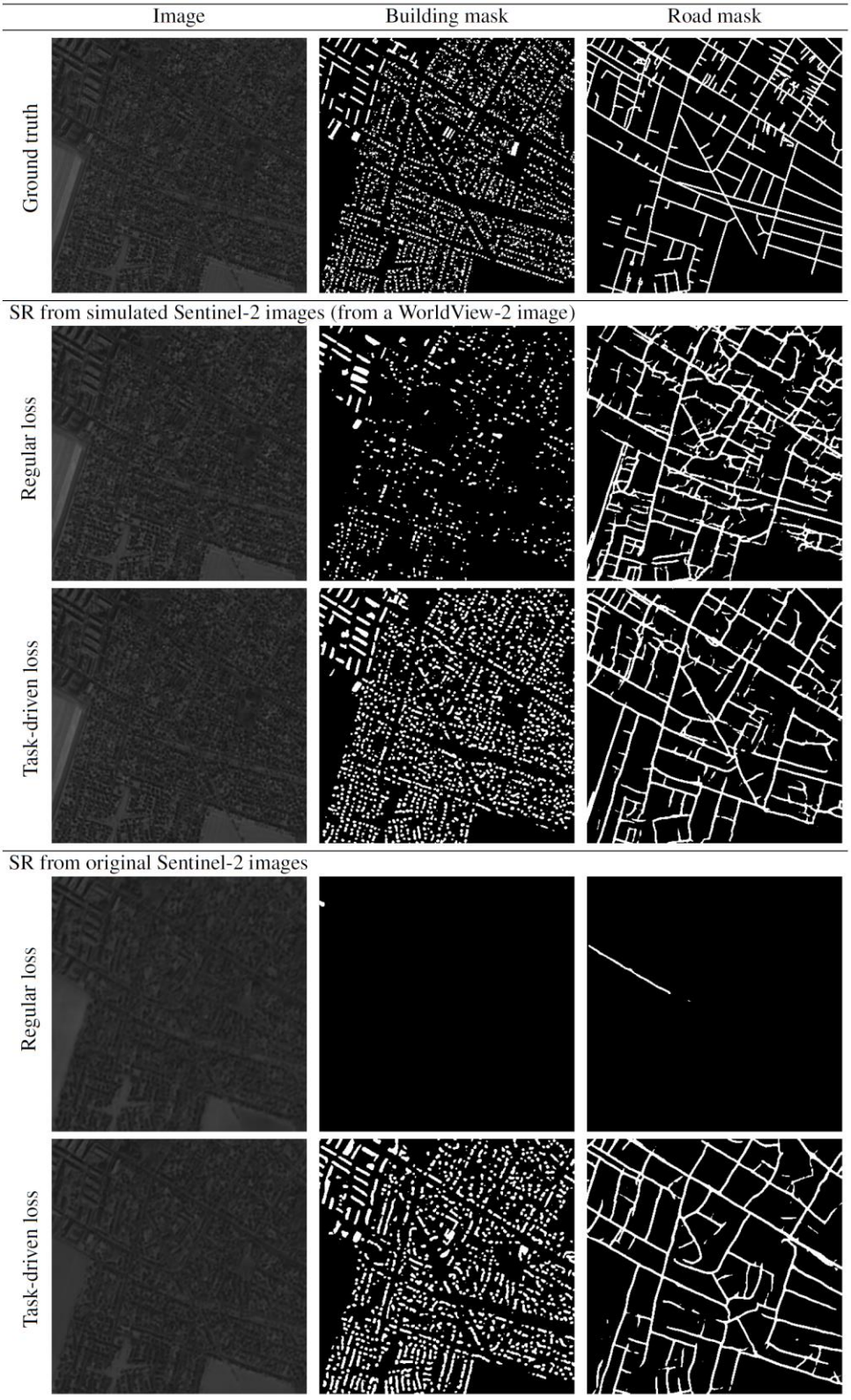


Figure 1: An example (an urban scene from the MuS2 benchmark) of reconstructing simulated (from WorldView-2) and real-world Sentinel-2 images using the RAMS network trained with regular and task-driven loss functions. The ground-truth image was acquired by WorldView-2 and the segmentation masks were extracted from OpenStreetMap.

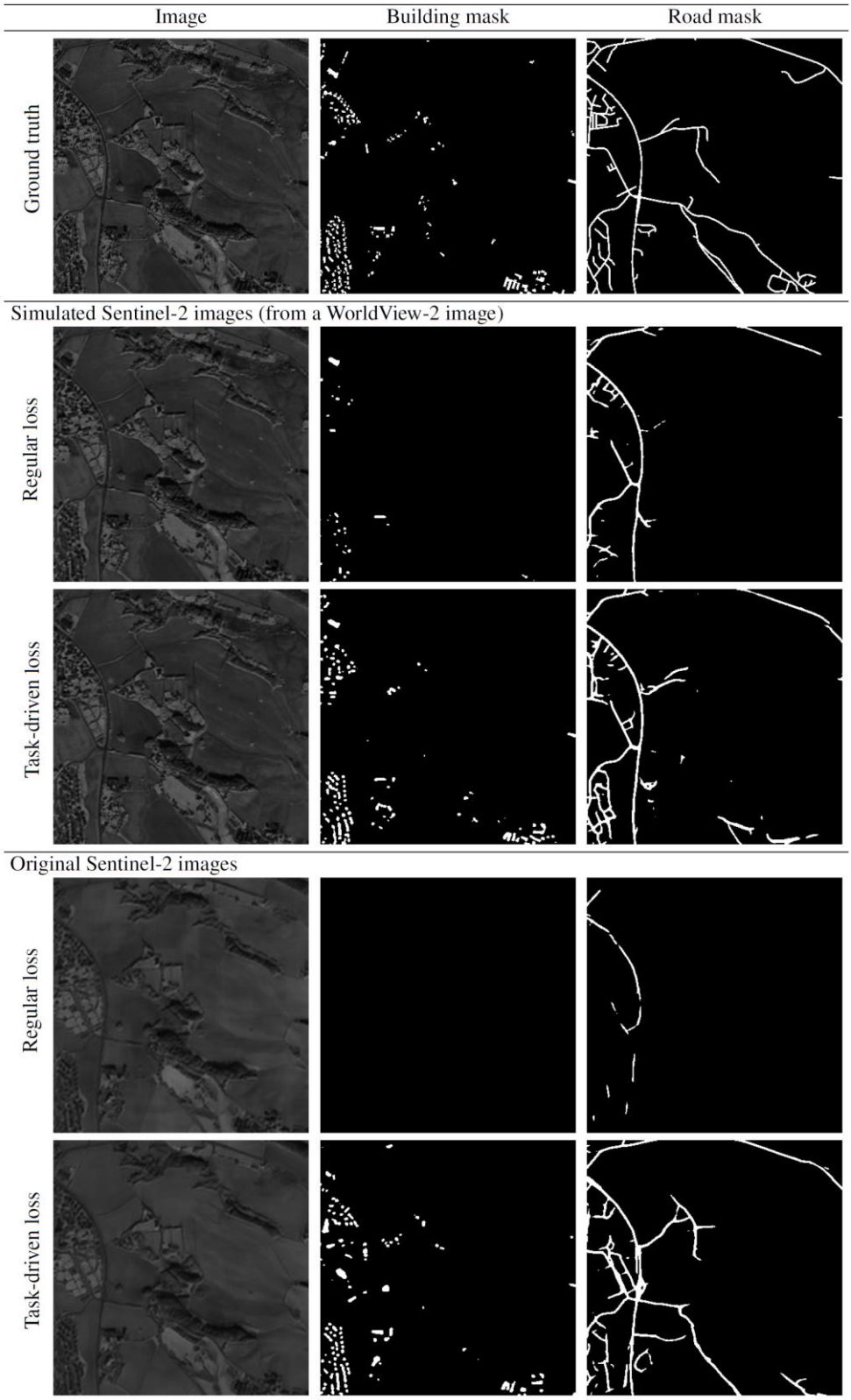


Figure 2: An example (a suburban scene from the MuS2 benchmark) of reconstructing simulated (fromWorldView-2) and real-world Sentinel-2 images using the RAMS network trained with regular and task-driven loss functions. The ground-truth image was acquired by WorldView-2 and the segmentation masks were extracted from OpenStreetMap.