Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2020 doi:10.20944/preprints202009.0678.v1

Technical Note

Satellite Image Multi-Frame Super-Resolution using 3D Wide-Activation Neural Networks

Francisco Dorr [1]

[1] Independent, José María Paz 745 12, Florida, 1602, Buenos Aires, Argentina; fdorr@dc.uba.ar

Abstract: The small satellite market continues to grow year after year. A compound annual growth rate of 17% is estimated during the period between 2020 and 2025. Low-cost satellites can send a vast amount of images to be post-processed at the ground to improve the quality and extract detailed information. In this domain lies the resolution enhancement task, where a low-resolution image is converted to a higher resolution automatically. Deep learning approaches to Super-Resolution (SR) reached the state-of-the-art in multiple benchmarks; however, most of them were studied in a single-frame fashion. With satellite imagery, multi-frame images can be obtained at different conditions giving the possibility to add more information per image and improve the final analysis. In this context, we developed and applied to the PROBA-V dataset of multi-frame satellite images a model that recently topped the European Space Agency's Multi-frame Super Resolution (MFSR) competition. The model is based on proven methods that worked on 2D images tweaked to work on 3D: the Wide Activation Super Resolution (WDSR) family. We show that with a simple 3D CNN residual architecture with WDSR blocks and a frame permutation technique as data augmentation better scores can be achieved than with more complex models. Moreover, the model requires few hardware resources, both for training and evaluation, so it can be applied directly from a personal laptop.

Keywords: multi-frame super resolution; wide activation super resolution; 3D convolutional neural network, deep learning

---

1. Introduction

In the past, the satellite market was reserved for a few companies and governments, which had the capacity (technical and monetary) to build and deploy large machinery in space, and the data obtained afterwards was used just by only a few research teams worldwide. Today, there is a growing interest, both social and commercial, in the deployment of small, low-cost satellites. A compound annual growth rate of 17% has been estimated for the small satellite market (forecast from 2020 to 2025) [1]. This expansion brings with it new challenges because of the vast amount of new data available. For example, satellite images are being used in many different fields to accomplish a wide spectrum of tasks. To name a few, Xu et al. [2] investigated vegetation growth trends over time, Martinez et al. [3] tracked tree growth through soil moisture monitoring, Ricker et al. [4] studied Arctic ice growth decay and Liu et al. [5] developed a technique to extract deep features from high-resolution images for scene classification.

But as satellites get more affordable and smaller, data quality cannot always be maintained; a trade-off must be found between price and quality. A case of study is high resolution (HR) images. They are not easy to obtain, or fast enough to transfer, and need costly and massive platforms as opposed to small, rapidly deployed, low-cost satellites that can provide viable services at the cost of lowering quality [6]. Image quality restrictions are common due to degradation and compression in the imaging process [7]. Notwithstanding, many tasks can be solved in post-processing steps, improving

 doi:10.20944/preprints202009.0678.v1

the quality once the data arrives on earth. Hence the importance of image resolution enhancement techniques that can take advantage of the huge and growing amount of information available from small satellites.

The problem of super-resolution (SR) is not new. It has been widely studied in different contexts taking multiple approaches. Specifically, to improve the resolution of satellite images, one common approach is the use of discrete wavelet transforms (DWTs), in which the input image is decomposed into different sub-bands and then combined to generate a new resolution through the use of inverse DWTs [8], [9], [10]. In recent years, as deep learning applications explode in every computer-vision task, convolutional neural networks methods began to dominate the problem of SR. However, most of them focus on single-image super resolution (SISR) [11], [12], [13] and do not take advantage of the temporal information inherent to multiframe tasks. MFSR has been studied in video; for example Sajjadi et al. [14] proposed a framework that uses the previously inferred HR estimate to super-resolve the subsequent frame, Jo et al. [15] created an end-to-end deep neural network that generates dynamic upsampling filters and a residual image avoiding explicit motion compensation, and Kim et al. [16] presented 3DSRnet, a framework that maintains the temporal depth of spatio-temporal feature maps to capture nonlinear characteristics between low and high resolution frames.

In this technical note we present a technique that takes as a strong baseline Kim et al. [16] 3DSRnet framework, but adapted for satellite image MFSR and replacing 3DCNN blocks with wide activation blocks [17]. This method core is a 3D wide activation residual network that was fully trained and tested on the PROBA-V dataset [18] on a low-specifications home laptop computer with only 4GB of GPU memory. Despite being low on resources and based on a simple architecture, this method topped Kelvin's ESA challenge in February 2020 [19].

2. Materials and Methods

2.1. Image data-set

In this work we used the set of images from the vegetation observation satellite PROBA-V of the European Space Agency (ESA) [18] provided in the context of the ESA's super resolution competition PROBA-V, which took place between 01.11.2018 and 31.05.2019 [20].

The PROBA-V sensors can cover 90% of the globe every day with a resolution of 300 meters (low resolution). Every 5 days they can provide images of 100 meters resolution (high resolution). With this in mind, the objective of the challenge is to build the 100m resolution images from multiple images of higher frequency of 300m resolution. It should be noted that the images provided for this challenge were not artificially degraded. As a common practice in super resolution developments, usually a high resolution image is artificially degraded and this is used as the low resolution starting point. In this case all the images are original, both the low and high resolution ones [20].

2.1.1. Data-set characteristics

As described in Märtens et al. [18], the data-set used for both training and testing is composed as follows:

- 1160 images from 74 hand-selected regions were collected at different points in time
- Divided in two spectral bands: RED with 594 images and NIR with 566. A radiometrically and geometrically corrected Top-of-Atmosphere reflectance in Plate Carre projection was used for both bands.
- LR size is 128 x 128, HR size is 384 x 384 and both have a bit-depth of 14 bits but saved as 16-bit png format.
- Each scene has a range of LR images (from a minimum of 9 to a maximum of 30) and one HR image.
- For each LR and HR image there is a mask that indicates which pixels can be reliably used for reconstruction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2020 doi:10.20944/preprints202009.0678.v1

2.2. Network architecture

The proposed 3DWDSRnet method for super resolution is based on a patch-based 3D-CNN architecture that allows multiple image inputs to be scaled into a single higher resolution image.

The problem being investigated is very similar to video SR, where the resolution of a single video frame is enhanced using the information from the surrounding frames. In a given sliding time window, video frames usually refer to a single scene but with subtle changes between each other. Thus, this temporal information can benefit resolution scaling more than single-image approaches [16]. The PROBA-V data set has multiple frames per location which can have shifts of up to one pixel. This evokes a similarity with the frames of a video and their possible variations and that is why we decided to investigate this path.

Our work takes as a strong baseline the framework proposed by Kim et al. [16] for video super-resolution: 3DSRnet. They use a 3D-CNN that takes five low-resolution input frames and seeks to increase the resolution of the middle frame. The network is a two-way residual network. The main path acts as a feature extractor from the chaining of multiple 3D convolutional layers that preserves the temporal depth. For the last layers, a depth reduction is performed to obtain the final 2D HR residual. Meanwhile, the second path takes the middle frame and applies a bicubic scaling. A Pixel Shuffle [21] reshapes the residual output of the main path, which is then added to the output of the secondary path obtained by this means the final HR frame (Figure 1).

Our approach differs from the original 3DSRnet in two main aspects (compare Figure 1 and 2):

- Convolutional layers are replaced by 3D WDSR blocks
- Bicubic upsampling is replaced by 2D WDSR blocks.

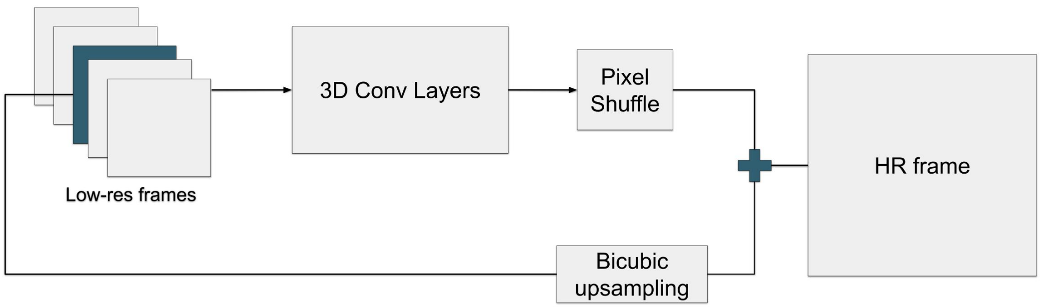

Figure 1. Original 3DSRnet: All low resolution input frames are fed into the main 3D Conv path to predict the residuals for the middle frame. The results of the paths are added up to obtain the final HR frame.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2020 doi:10.20944/preprints202009.0678.v1

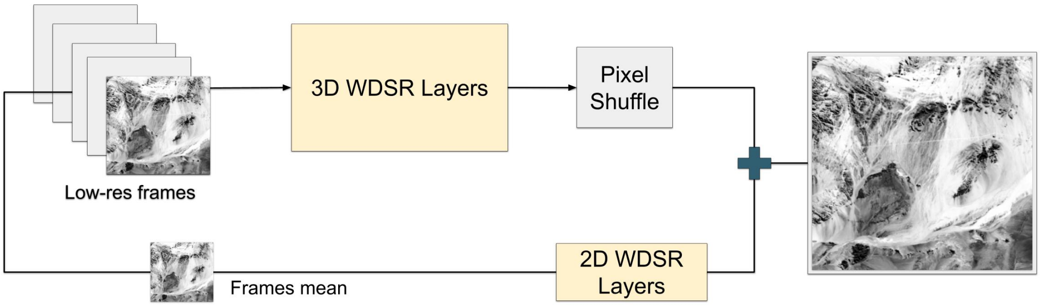

Figure 2. 3DWDSRnet: All low-resolution input frames are fed into the main WDSR Conv 3D path to predict the scene's residuals.The average of the frames is used as input to the WDSR 2D Convs path. The results of both paths are added together to obtain the final RH frame. Soft yellow highlights the differences with the original 3DSRnet blocks.

2.2.1. WDSR blocks

Yu et al. [17] describe WDSR blocks as residual blocks with the capability to increase the final accuracy of a SISR task. They demonstrate that a feature expansion using a 1x1 Conv before the ReLU activation, followed by a feature factorization given by a 1x1 Conv and a 3x3 Conv keeps more information and even lowers the number of total parameters used (Figure 3). This residual block was named WDSR-b. In our study we expand the notion of WDSR-b block from 2D to 3D and replace every single 3D Conv from 3DSRnet with it. To do so, we simply change kernel sizes from 1x1 to 1x1x1 and from 3x3 to 3x3x3. Everything else remains exactly the same. The implementation of the WDSR block was based on Krasser's github code [22].

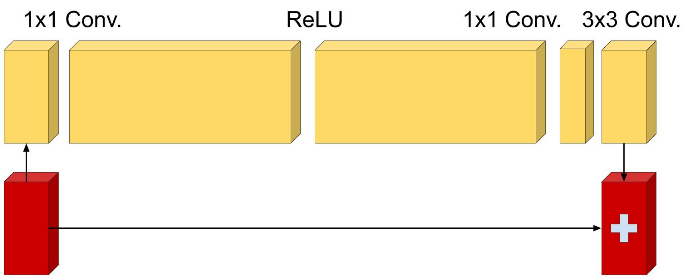

Figure 3. 2D-WDSR-b block. The residual block is composed of a 1x1 Conv to expand features before ReLU activation. After activation a 1x1 Conv followed by a 3x3 Conv are applied. 3D-WDSR-b used by our method follows the same approach but taking into account the time dimension.

2.3. Preprocessing and Data Augmentation

The preprocessing steps were performed as follows:

1. Register all frames from each image to the corresponding first frame using masked_registered_translation function from scikit-image [23].
2. Remove images where all of their frames had more than 15% dirty pixels.
3. Select K best frames (from cleanest to dirtiest, k=7).
4. Extract 16 patches per image.
5. Remove instances where the HR target patch had more than 15% dirty pixels.

To make the training more robust to the pixel shifts and differences between frames, a frame-basis data-augmentation was performed. For each patch, 6 new patches were added to the training set. Each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2020 doi:10.20944/preprints202009.0678.v1

of them with a random frame permutations. A similar augmentation technique can be found in Rifat et al. [24]. The impact of patches and data augmentation by frame permutation can be seen in Section 3.

2.4. Training

The training was performed on a low-end laptop GPU GTX1050 with 4GB of memory. First, the model was trained on NIR band until no more improvements were found (156 epochs). Then, RED band was trained over the NIR band pretrained model (61 epochs). This two-step model training was based on Molini et al. [25] where they found it increased the final accuracy.

We used NAdam optimizer with a learning rate of 5e-5, patch size of 34 with a stride of 32, and a batch size of 32 patches. The main path was composed of 8 3DWDSR-b blocks before the time dimension reduction. As regularization it is important to note that common techniques such as Batch Normalization do not work well in SR problems ([26], [27]). We used Weight Normalization as recommended by Yu et al. [17].

2.4.1. Quality metric and loss function

Märtes et al. [18] proposed the quality metric clean Peak Signal-to-Noise Ratio (cPSNR) that takes into account the pixel shifts between frames and is applicable for images which only have partial information (dirty pixels). Basically, cPSNR ignores masked pixels (due to wrong pixels, clouds, etc.) and takes into account all possible pixel shifts within frames to calculate the final metric. Inspired by cPSRN, Molini et al. [25] propose cMSE, a modified mean square error to use as loss functions. Since SR prediction and HR target could be shifted, the loss embeds a shift correction. To do so, SR is cropped at the center by d pixels (defined by the maximum expected shift between images) and all possible patches shifts $HR_{u,v}$ are extracted from the target HR image. Thereafter, all possible MSE scores are calculated for each $HR_{u,v}$ patch and the minimum score is taken.

We found that this loss works quite well for the problem but based on Zhao et al. [28] we follow their recommendation to use Mean Absolute Error loss (MAE) as a substitute. Mathematically, cMAE is defined as follows:

$$L = \min_{u,v\in[0,2d]} \frac{\sum_{i=1}^{N_{u,v}} |HR_{u,v} - (SR_{crop} + b)|}{N_{u,v}}$$

where $N_{u,v}$ is the total number of clean pixels in u, v crop and b is the brightness bias corrections. Figure 4 shows how the loss is calculated taking into account all possible pixel shifts.

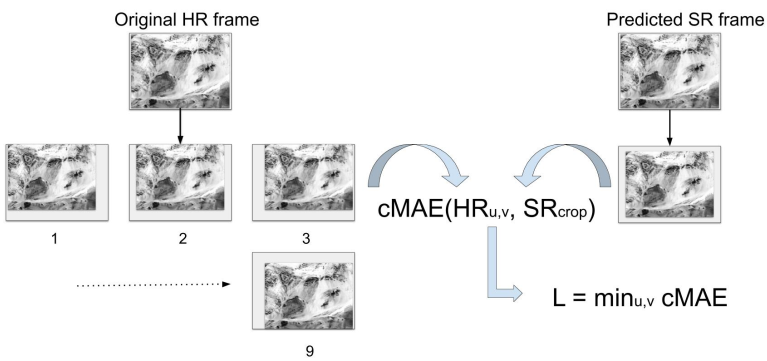

Figure 4. cMAE. For each possible pixel shift u,v the cMAE is calculated between cropped SR and $HR_{u,v}$ patch. In this example figure, the maximum possible shift is 1 (both horizontally and vertically), so 9 patches are extracted for each possible combination. PROBA-V dataset has a maximum of 3 pixel shift, so 49 combinations are needed to calculate the final loss.

3. Results

Table 1. Scores obtained in the Kelvins ESA's competition public leaderboard (Feb 2020). Scores are normalized by baseline, less is better.

Method	Patch	Frames	Loss	Normalization	Score	Memory requirement
DeepSUM	96x96 (bicubic)	9	cMSE	Instance	0.94745	+++
HighResnet	64x64	16	cMSE	Batch	0.94774	++
3DWDSRnet (ours)	34x34	5	cMAE	Weight	0.97933	+
3DWDSRnet (ours)	34x34 (aug)	5	cMAE	Weight	0.96422	+
3DWDSRnet (ours)	34x34 (aug)	7	cMAE	Weight	**0.94625**	+

3.1. Comparisons

We compare 3DWDSRnet to top methods at the moment the investigation was performed (February 2020): DeepSUM [25] and HighResnet [24] (1) . It is worth noting that Kelvin ESA Competition is still open to teams that want to try their solution in a post-mortem leaderboard. As the code of 3DWDSRnet is open to use, modify and share, a github repository [29] was provided. There are teams that at the time of writing this technical note has built upon it and are still improving the metrics [30].

Description of compared methods in Table 1:

- **DeepSUM**: it performs a bicubic upsampling of images before feeding the network. When using this approach higher specifications are needed because of increasing memory cost making it impossible to train in a low-specifications equipment.
- **Highres-net**: this method upscales after fusion, so memory usage is reduced. However, they still need 16 frames and 64x64 patches to reach the best score. Scores are improved by averaging the outputs of two pretrained networks.
- **3DWDSRnet**: our method follows Highres-net approach of upscaling after fusion but achieves similar scores using less than half of the image frames (7) and half size patch size (34x34). Moreover, there is no need of averaging two methods to obtain these results.

4. Discussion

The results address some interesting insights about the common methods used in MFSR. It is shown that not always the more complex and memory consumption architectures are indeed the best ones. Sometimes a simple model but with the correct parameters performs better. For example, in our method, increasing frames from 5 to 7 shows an increased performance (1). Moreover, tweaking the data outside the neural network can improve the metrics even more. A simple method such as frame permutation for data augmentation shows a consistent growth in the score.

This makes us wonder if the money and work hours invested in designing the architecture of new neural networks as a general problem-solving approach is always a good choice. We point out, instead, the need to develop, in addition, algorithms optimized for every need. The latter could be used by teams or individuals with less hardware resources. The access of third world countries to the latest advances in hardware is not always possible, so in order to democratize the access to AI worldwide, more research should be done on accessible but equally efficient models.

This technical note serves as a base to continue improving the 3DWDSRnet method. Some possible directions to explore are:

- Further investigate data augmentation methods to take benefits of multiple frames such as more interesting permutations, insert of pixel variations simulating clouds, and changes in image color, brightness, contrast, etc.
- Ensemble results from multiple models as done in Highres-net [24].
- Try different patch sizes and see how this affects the performance.

Funding: This research received no external funding.

Acknowledgments: Joaquín Padilla Montani, Fernando G. Wirtz and Ricardo A. Dorr for general corrections.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SR	Super Resolution
MISR	Multi-image Super Resolution
MFSR	Multi-frame Super Resolution
CNN	Convolutional Neural Network
Conv	Convolutional
AI	Artifitial Intelligence
MAE	Mean Absolute Error
MSE	Mean Square Error
cPSNR	clean Peak Signal-to-Noise Ratio

References

1. Knuth, M.I. Small Satellite Market - Growth, Trends, and Forecast (2020 - 2025). Available online: https://www.mordorintelligence.com/industry-reports/small-satellite-market (accessed on 23 September 2020).
2. Xu, H.j.; Wang, X.p.; Yang, T.b. Trend shifts in satellite-derived vegetation growth in Central Eurasia, 1982–2013. *Science of the Total Environment* **2017**, *579*, 1658–1674.
3. Martínez-Fernández, J.; Almendra-Martín, L.; de Luis, M.; González-Zamora, A.; Herrero-Jiménez, C. Tracking tree growth through satellite soil moisture monitoring: A case study of Pinus halepensis in Spain. *Remote Sensing of Environment* **2019**, *235*, 111422.
4. Ricker, R.; Hendricks, S.; Girard-Ardhuin, F.; Kaleschke, L.; Lique, C.; Tian-Kunze, X.; Nicolaus, M.; Krumpen, T. Satellite-observed drop of Arctic sea ice growth in winter 2015–2016. *Geophysical Research Letters* **2017**, *44*, 3236–3245.
5. Liu, Q.; Hang, R.; Song, H.; Li, Z. Learning multiscale deep features for high-resolution satellite image scene classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *56*, 117–126.
6. Sweeting, M. Modern small satellites-changing the economics of space. *Proceedings of the IEEE* **2018**, *106*, 343–361.
7. Luo, Y.; Zhou, L.; Wang, S.; Wang, Z. Video satellite imagery super resolution via convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters* **2017**, *14*, 2398–2402.
8. Demirel, H.; Anbarjafari, G. Satellite image resolution enhancement using complex wavelet transform. *IEEE geoscience and remote sensing letters* **2009**, *7*, 123–126.
9. Demirel, H.; Anbarjafari, G. Discrete wavelet transform-based satellite image resolution enhancement. *IEEE transactions on geoscience and remote sensing* **2011**, *49*, 1997–2004.
10. Iqbal, M.Z.; Ghafoor, A.; Siddiqui, A.M. Satellite image resolution enhancement using dual-tree complex wavelet transform and nonlocal means. *IEEE geoscience and remote sensing letters* **2012**, *10*, 451–455.
11. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2472–2481.
12. Sun, W.; Chen, Z. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing* **2020**, *29*, 4027–4040.
13. Anwar, S.; Barnes, N. Densely residual laplacian super-resolution. *arXiv preprint arXiv:1906.12021* **2019**.
14. Sajjadi, M.S.; Vemulapalli, R.; Brown, M. Frame-recurrent video super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6626–6634.
15. Jo, Y.; Wug Oh, S.; Kang, J.; Joo Kim, S. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3224–3232.

doi:10.20944/preprints202009.0678.v1

16. Kim, S.Y.; Lim, J.; Na, T.; Kim, M. 3DSRnet: Video Super-resolution using 3D Convolutional Neural Networks. *arXiv preprint arXiv:1812.09079* **2018**.
17. Yu, J.; Fan, Y.; Yang, J.; Xu, N.; Wang, Z.; Wang, X.; Huang, T. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718* **2018**.
18. Märtens, M.; Izzo, D.; Krzic, A.; Cox, D. Super-resolution of PROBA-V images using convolutional neural networks. *Astrodynamics* **2019**, *3*, 387–402.
19. Website, K.E.A.C.C. PROBA-V Super Resolution post mortem. Available online: https://kelvins.esa.int/proba-v-super-resolution-post-mortem/leaderboard (accessed on 23 September 2020).
20. Website, K.E.A.C.C. PROBA-V Super Resolution Competition. Available online: https://kelvins.esa.int/proba-v-super-resolution (accessed on 23 September 2020).
21. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.
22. krasserm. Single Image Super-Resolution with EDSR, WDSR and SRGAN. Available online: https://github.com/krasserm/super-resolution (accessed on 23 September 2020).
23. Van der Walt, S.; Schönberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T. scikit-image: image processing in Python. *PeerJ* **2014**, *2*, e453.
24. Rifat Arefin, M.; Michalski, V.; St-Charles, P.L.; Kalaitzis, A.; Kim, S.; Kahou, S.E.; Bengio, Y. Multi-Image Super-Resolution for Remote Sensing Using Deep Recurrent Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 206–207.
25. Molini, A.B.; Valsesia, D.; Fracastoro, G.; Magli, E. DeepSUM: Deep neural network for Super-resolution of Unregistered Multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *58*, 3644–3656.
26. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136–144.
27. Fan, Y.; Shi, H.; Yu, J.; Liu, D.; Han, W.; Yu, H.; Wang, Z.; Wang, X.; Huang, T.S. Balanced two-stage residual networks for image super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 161–168.
28. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861* **2015**.
29. Dorr, F. 3DWDSR: Multiframe Super Resolution Framework applied to PROBA-V challenge. Available online: https://github.com/frandorr/PROBA-V-3DWDSR (accessed on 23 September 2020)., 2020. doi:10.5281/zenodo.3634101.
30. Bajo, M. Multi-Frame Super Resolution of unregistered temporal images using WDSR nets. Available online: https://github.com/mmbajo/PROBA-V (accessed on 23 September 2020)., 2020. doi:10.5281/zenodo.3733116.

Sample Availability: Trained models and code public repository: https://github.com/frandorr/PROBA-V-3DWDSR