Degradation-Aware Self-Supervised Multi-Temporal Super-Resolution

DEGRADATION-AWARE SELF-SUPERVISED MULTI-TEMPORAL SUPER-RESOLUTION

Matteo Impieri, Diego Valsesia, Tiziano Bianchi, Enrico Magli

Politecnico di Torino
Department of Electronics and Telecommunications
Torino, Italy

ABSTRACT

Learning deep super-resolution models without the need for ground truth data at a higher resolution is critical for satellite imaging applications. This is either due to the lack of existing images at better resolution for certain target wavelengths or the existence of significant domain gaps between the images of different satellites. In this paper, we propose a method and neural network architecture for a multi-image super-resolution problem, where each image in the input stack might be affected by a different degradation process. A test-time finetuning procedure allows to dynamically account for the degradations observed for a specific set of LR inputs, improving over baseline results.

***Index Terms*—** Multi-image super-resolution, kernel estimation, self-supervised learning.

1. INTRODUCTION

Numerous remote sensing applications demand the acquisition of highly detailed images, but this is often hindered by limitations in satellite sensor capabilities and communication channels. Recently, multi-image super-resolution (MISR) techniques showed that powerful deep-learning models can effectively combine multiple low-resolution (LR) images of the same location at different times to reconstruct a high-resolution image (HR), overcoming challenges like varying illumination, cloud cover, and temporal changes.

However, existing literature predominantly approaches the MISR problem from a supervised learning perspective [1, 2, 3, 4], relying on ground truth HR images during training. This is a common setup, for example, with the Proba-V challenge dataset [5] where ground truths at higher resolution are available due to the unique nature of the Proba-V setting. This poses challenges, as collecting data at higher resolutions can be expensive or not feasible at all. In fact, for certain wavelengths there might not be a satellite in operation that already provides better imagery. In other cases, there might be a significant mismatch between the features of images captured by two different satellites, causing a domain gap that hinders model learning and successful inference.

For these reasons, more research is needed in the direction of methods that do not require imagery at higher resolution. This setting is particularly challenging for MISR techniques where a variety of factors might affect the degradation that each of the images in the time series has undergone. In our previous investigation [6], we showed the potential of unsupervised deep learning methods to improve over classic techniques, even for the MISR setting, while still reporting a gap with respect to the performance achievable by supervised training. Moreover, our previous work did not provide an end-to-end system, relying instead of handcrafted degradation kernels for the training process.

In this paper, we employ data-driven techniques to extract estimates of real degradation kernels from the available LR images and train a state-of-the-art MISR neural network in a self-supervised manner. The training process is such that the network learns to be robust to a wide set of realistic degradations, as well as to the fact that each image in the LR input time series might be affected by a different degradation. Moreover, we show that a test-time finetuning procedure is able to dynamically calibrate the pretrained neural network to the specific degradation present in the test input, yielding superior results.

2. BACKGROUND

The challenge of image Super-Resolution (SR) has garnered significant attention over the years, particularly witnessing notable advancements recently due to the application of deep learning methods. Existing literature, encompassing both conventional photographs and remote sensing images, has predominantly concentrated on Single-Image SR (SISR). Common approaches involve supervised training, where High-Resolution (HR) images at the target resolution are essential, following either paired or unpaired methods.

The literature on blind SISR [7] is particularly relevant to this work, emphasizing the importance of understanding

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

the degradation process that generates LR observations from HR images in real-world SR. This is key to enable training that does not require ground truth images at the target resolution. Unsupervised or self-supervised training often relies on assuming invariance across scales [8], where the function mapping LR to HR images is assumed to be the same for any LR-HR resolution pair affected by the same degradation process. This hypothesis means that is is possible to generate further degraded images at a coarser resolution (CR) directly from LR observations and to train a model to recover the original LR. Post-training, the model is applied in an extrapolation mode to map the LR image to a higher resolution, but the success depends on accurate modeling of the degradation process; any mismatch significantly degrades SR performance.

For what concerns the MISR setting, most of the current attention has been on supervised training [2, 3], with only a few exceptions [9]. In remote sensing, the Proba-V dataset has been instrumental for many MISR works due to LR and HR images at different resolutions from the same platform. Emerging datasets [10] offer increased diversity and higher resolution imagery, potentially benefiting the further development of both supervised and unsupervised methods. In our earlier work [6], we explored the potential of unsupervised training by handcrafting some degradation models in order to train the PIUNet model [1] in a self-supervised manner. The work showed promising performance, although the gap with respect to supervised training suggested that more work was needed, e.g. in the direction of a fully data-driven method.

3. METHOD

In this section, we explore the design of a MISR technique based on deep neural networks that can be trained from LR images only and is aware of the degradation operator that generated each of the input images. In order to train the model with only LR images, we resort to the commonly-used self-supervised approach that relies on scale invariance. In practice, the LR images are further degraded with operators that resemble the ones that actually generated the LR images and downsampled to a coarse resolution (CR). The model then is trained to recover an LR image from a set of CR images.

3.1. Self-supervised training

Since we are dealing with a MISR setting, rather than SISR, it is important to remember that the multiple LR images that are provided as input may be affected by different degradations. In fact, the degradation is not only a function of the optics but also of the specific processing steps, such as orthorectification, used on it, thus making the degradation operator time-varying and, possibly, space-varying. The first step towards self-supervised training is therefore estimating a set of degradation operators to be used for the LR-to-CR degradation, such that they are as faithful as possible to the real HR-to-LR degradations. In our earlier exploration [6], we handcrafted such operators to demonstrate the potential of unsupervised MISR. In this work, we use a data-driven approach to directly estimate realistic degradation operators from the LR images themselves.

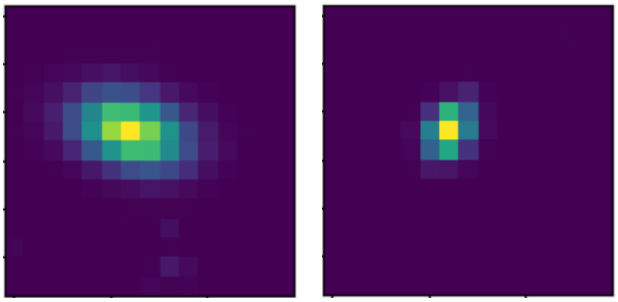

Fig. 1. Examples of degradation kernels from LR Proba-V images as estimated by DIP-FKP.

To this end, we use the DIP-FKP [11] approach to kernel estimation on each of the LR images. The approach uses two neural networks, an image generator (DIP) and a kernel generator (FKP), that can be optimized on a single LR image in order to estimate the kernel of the degradation filter used from HR to LR. The image generator maps a latent noise $\mathbf{z}_x$ to an HR image, while the kernel generator maps a different latent noise $\mathbf{z}_k$ to a filter kernel. The generated HR image is then filtered with such kernel and downsampled to generate an LR image to be compared with the available LR image. The kernel generator is pretrained to generate anisotropic Gaussian kernels, while DIP is randomly initialized. In order to find the degradation kernel associated to a specific LR image, an optimization problem is solved to minimize the error between the generated LR image and the target one with respect to the image generator weights and the kernel latent noise $\mathbf{z}_k$. Once this operation is repeated for all the LR images in the training set, we have a collection of filter kernels that approximate the degradations that generated the LR images. Examples of these kernels are shown in Fig. 1.

The estimated kernels are then randomly selected to degrade T LR images of a multi-temporal series to CR, which will serve as input to the PIUnet neural network model for MISR. PIUnet is trained with the L1 loss between its output and one of the LR images that generated the CR inputs. This procedure makes PIUnet more robust to realistic degradation kernels which may vary from image to image.

3.2. Test-time finetuning

The previously explained approach follows a conventional self-supervised approach to building robustness to kernel variations. However, in learning a general super-resolution model, it does not exploit knowledge of the specific degradation affecting a specific input. This could enhance the

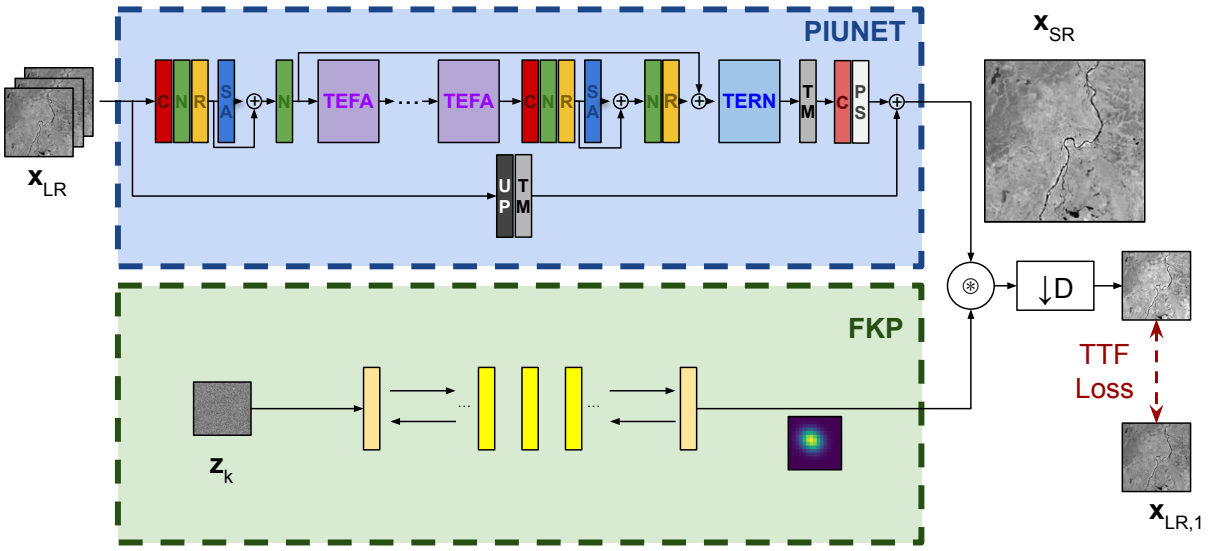

Fig. 2. Test-time finetuning (TTF) of PIUnet to calibrate for the kernel extracted by FKP for the specific LR input.

performance of a model pretrained in the previous manner, by acting as a sort of "calibration" of the neural network to the specific degradations observed for a particular input. In order to achieve this we propose to use a mechanism similar to the DIP-FKP approach used in kernel estimation but for this calibration purpose. An overview is shown in Fig. 2. In particular, the DIP image generator is replaced by the pretrained PIUnet model, while leaving the FKP branch as is. At test time, for a given input of T LR images $\{\mathbf{x}_t^{\text{LR}}\}_{t=1}^{T}$, the following optimization problem is solved:

$$\min_{\boldsymbol{\Theta},\mathbf{z}_k} \left\| \left[F_{\boldsymbol{\Theta}}(\{\mathbf{x}_t^{\text{LR}}\}_{t=1}^{T}) * G(\mathbf{z}_k) \right]_{\downarrow D} - \mathbf{x}_1^{\text{LR}} \right\|_1 \quad (1)$$

being $F_{\boldsymbol{\Theta}}$ the PIUnet MISR network parameterized by weights $\boldsymbol{\Theta}$, G the FKP kernel generator and $*$ 2D convolution. Notice how this optimization problem seeks to i) optimize the latent noise $\mathbf{z}_k$ that when passed though the FKP network generates the kernel that is most likely to model the HR-to-LR degradation; ii) finetune the PIUnet weights. This finetuning acts as a calibration of PIUnet for the specific input subject to the degradation kernel estimated by FKP. We remark that over-optimization of the objective in Eq. (1) leads to overfitting but locally acts as a calibration, in the same vein as deep image prior [12] locally acts as a prior.

4. EXPERIMENTAL RESULTS

In this section, we present some experimental results to benchmark the proposed self-supervised method against supervised approaches and the non-data-driven solution in [6] in a fair setting with equal neural network architectures. For this benchmark, we use the Proba-V dataset for both training and testing. Self-supervised training only uses the LR data, but the availability of real HR data for testing allows us to quantitatively measure SR performance in a real setting without having the degradation process under our control. The paired nature of the dataset also allows straightforward comparisons with supervised training using the HR data. The PIUnet archicture is also used for the supervised setting, while we also report the performance of the classic method of iterative backprojection (IBP) [13]. All methods use $T = 9$ input LR images to produce a single SR image. Concerning the proposed method, the FKP module has been pretrained to generate kernels of size 15×15. The test-time finetuning has been run for a total of 50 iterations for each test scene.

Table 1 report the cPSNR metric [5] on the Proba-V test set for the various methods under consideration. It can be noticed that the results obtained with only the self-supervised training on the the estimated kernels yields similar results to the results with handcrafted kernels reported in [6]. However, the calibration procedure proposed for test-time finetuning (TTF) is able to substantially improve performance.

We remark that a limitation of the proposed method is the estimation of shift-invariant kernels. Indeed, as previously discussed in [6], and emphasised by the results in Table 1, a more suitable model would be to have spatially-varying per-pixel kernels to properly account for effects of processing like orthorectification. In this regard, we report a negative result obtained during the development of this work, in which a state-of-the-art per-pixel kernel estimation network [14] failed to generate physically-plausible kernels from Proba-V images. Further work is therefore needed to extend the current approach to spatially-varying degradation models.

Table 1. Quantitative performance - cPSNR (dB)

	Classic	**Deep unsupervised**				**Deep supervised**
	IBP [13]		PIUnet		**PIUnet-FKP TTF**	PIUnet [1]
	-	Handcrafted [6]	Handcrafted [6]	**Learned kernel**	**Learned kernel**	-
	-	shift-invariant kernel	per-pixel kernel	**with DIP-FKP**	**with DIP-FKP**	-
NIR	45.96	46.78	46.98	46.84	47.06	48.41
RED	48.21	49.02	48.97	48.86	49.25	50.53

5. CONCLUSIONS

We presented an approach to self-supervised multi-image super-resolution that is capable of estimating real degradation kernels for the available LR images and use them to train a MISR neural network that is robust to multiple degradations and does not require ground truth images at better resolution. We also showed a test-time finetuning procedure that is capable of dynamically estimating the specific degradations of the current input images and calibrate the pretrained MISR network accordingly.

6. REFERENCES

[1] Diego Valsesia and Enrico Magli, "Permutation invariance and uncertainty in multitemporal image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[2] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli, "Deepsum: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2019.

[3] Francesco Salvetti, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sensing*, vol. 12, no. 14, 2020.

[4] Md Rifat Arefin, Vincent Michalski, Pierre-Luc St-Charles, Alfredo Kalaitzis, Sookyung Kim, Samira Ebrahimi Kahou, and Yoshua Bengio, "Multi-image super-resolution for remote sensing using deep recurrent networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. 2020, pp. 816–825, IEEE.

[5] Marcus Märtens, Dario Izzo, Andrej Krzic, and Daniël Cox, "Super-resolution of proba-v images using convolutional neural networks," *Astrodynamics*, vol. 3, no. 4, pp. 387–402, 2019.

[6] Nicola Prette, Diego Valsesia, Tiziano Bianchi, and Enrico Magli, "Towards unsupervised multi-temporal satellite image super-resolution," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 5135–5138.

[7] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong, "Blind image super-resolution: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[8] Assaf Shocher, Nadav Cohen, and Michal Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3118–3126.

[9] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo, "L1bsr: Exploiting detector overlap for self-supervised single-image super-resolution of sentinel-2 l1b imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2012–2022.

[10] Julien Cornebise, Ivan Orsolic, and Freddie Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset – with application to super-resolution," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[11] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte, "Flow-based kernel prior with application to blind super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10601–10610.

[12] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.

[13] Michal Irani and Shmuel Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[14] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Mutual affine network for spatially variant kernel estimation in blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4096–4105.