Towards Unsupervised Multi-Temporal Satellite Image Super-Resolution

TOWARDS UNSUPERVISED MULTI-TEMPORAL SATELLITE IMAGE SUPER-RESOLUTION

Nicola Prette, Diego Valsesia, Tiziano Bianchi, Enrico Magli

Politecnico di Torino - Torino, Italy

ABSTRACT

Multi-temporal super-resolution (SR) whereby a number of images of the same scene acquired at different times are fused to enhance its spatial resolution has recently enjoyed great success thanks to advances in deep learning methods. However, the literature has so far focused on supervised training approaches that require the availability of high-resolution (HR) images at the target resolution. This is a significant limitation because such imagery may not exist, might be difficult to source or exhibit domain gaps such as different spectral bands or radiometric characteristics. Unsupervised training approaches that do not require imagery beyond the input low resolution are needed to overcome this limitation. This paper presents a first analysis of the problem, taking inspiration from the literature on blind single-image SR, but also focusing on the uniqueness of multi-temporal satellite images. Our preliminary results show that it is indeed possible to develop accurate deep learning models for multi-temporal SR without HR images.

Index Terms— Super-resolution, multitemporal, Proba-V, unsupervised.

1. INTRODUCTION

Several applications in the field of remote sensing require the capture of very high resolution images. However, this is often not possible due to limitations in the capabilities of the sensors employed on-board of satellites, and in the channel capacity between the satellite and Earth. Multi-image super-resolution (MISR) techniques have recently enjoyed great success in addressing such scenarios. In particular, the ESA Proba-V challenge [1] stimulated research on powerful deep-learning models that can effectively fuse multiple low-resolution (LR) images of the same location at different times to reconstruct a single high-resolution image (HR), overcoming challenges like differences in illumination, cloud coverage and temporal change in the scene.

Still, current literature [2, 3, 4] is focused on the MISR problem from a supervised learning perspective, which means that ground truth HR images are available to the training process. This is problematic, because, except for a few cases, like Proba-V, where LR and HR images are taken from the same platform, it requires collecting data from multiple satellites. This causes domain gaps since multiple satellites will have mismatched radiometric properties, causing artefacts or suboptimal results. It is also worth noting that HR products may be difficult to obtain, especially at certain wavelengths, for very high resolutions, or in the development of new instruments, further limiting the applicability of supervised training.

In this paper, we take the first steps towards addressing the unsupervised MISR problem, where our objective is to develop powerful deep-learning-based SR models from LR images only to overcome the data requirements and features mismatch of current supervised training processes. We notice that a significant body of work in this direction is emerging for the single-image SR problem [5], but the multi-image problem in the remote sensing context is yet unexplored. We show that careful consideration should be placed on modelling the degradation process that results in the multi-temporal LR observations. We present preliminary results using the Proba-V dataset which can provide a reliable HR ground truth for performance assessment. Our results, obtained with handcrafted degradation models, show that unsupervised training is indeed possible, although it incurs a quality penalty with respect to supervised training. Bridging this final gap will be the focus of future work on more advanced deep learning models including degradation estimation.

2. BACKGROUND

Image SR is a problem that has received great attention over the years and has recently enjoyed significant improvements thanks to deep learning methods. Most of the literature [6], regarding both conventional photographs and remote sensing images, has focused on single-image SR (SISR). Typical approaches involve supervised training, which relies on availability of HR images at the target resolution, following either

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.

paired or unpaired approaches. We remark that some authors [7] use the "unsupervised" term to describe works using unpaired data, i.e., where HR images are available but not from the same scenes as LR images. This marginally alleviates data requirements, but does not solve the intrinsic difficulty of needing images at the target high resolution. In this paper, we use the "unsupervised" term in a stricter matter, i.e., to indicate that *no images at resolution beyond the LR observations are needed for training*. Of particular interest to the work in this paper is the literature on blind SISR [5], where it is shown that knowledge of the degradation process that generates LR observations from HR images is critical for real-world SR. In fact, one of the most commonly used approaches to unsupervised training of SR models is to assume invariance across scales, i.e., the function to be learned to map from LR to HR images is the same for any pair of "LR" or "HR" resolutions coupled by the same degradation process. Under this hypothesis, unsupervised training becomes self-supervised as it is possible to generate degraded images at a coarser resolution (CR) directly from LR observations and train a model to recover the original LR. After training, one uses the model in an extrapolation regime to map the LR image to a higher resolution. The blind SR literature shows that this works as long as the degradation process faithfully models that which generated the LR observations and any mismatch will significantly degrade SR performance.

MISR has so far largely focused on supervised training. In the computational photography setting, focus has been placed on the burst SR [8] problem where a camera (typically a smartphone) acquires a set of photos in rapid succession with possibly camera and scene movements in between captures. Recent burst SR techniques have been developed under supervised conditions where an HR ground truth of the scene is captured by a separate camera, typically with a telephoto lens [9]. In remote sensing, the Proba-V dataset has been used for numerous MISR works thanks to the availability of both LR images at 100m resolution and HR images at 300m from the same platform and with a non-trivial degradation function connecting the two acquisitions. New datasets [10] are emerging with increased diversity and with higher resolution imagery, which could be useful for further development of both supervised and unsupervised methods.

3. MODEL AND BASELINE APPROACH

The goal of this section is to study how to correctly frame the problem of unsupervised MISR, discuss the main ingredients that are needed for future development of successful models, and develop some insights on the problem, specifically for remote sensing imagery. Subsequently, we will provide a baseline of what unsupervised approaches could achieve and how they compare with respect to the state-of-the-art in the supervised literature.

A baseline approach [11] towards unsupervised SR is to consider the SR function as approximately scale-invariant. This allows us to synthetically degrade the LR images to a coarser resolution (CR) using a degradation model and train a neural network to recover the original LR image. When testing on real images, we assume that the learned function can extrapolate to the generation of a higher resolution image from the LR input. For this scope, correct modeling of the degradation kernel is pivotal as it is known that training on the wrong kernel leads to poor generalization [12]. We argue that the following degradation model is suitable for a multitemporal set of T satellite images:

$$\mathbf{X}_{\mathrm{LR},t} = [\mathbf{K}_t \mathbf{X}_{\mathrm{HR}}]_{\downarrow D} + \mathbf{n}_t \qquad t = 1, \ldots, T \tag{1}$$

where $\mathbf{K}_t$ represents a spatially-variant and time-dependent degradation kernel, $\downarrow D$ the decimation operator by a factor of D, and $\mathbf{n}$ represents additive noise. The application of kernel $\mathbf{K}_t$ amounts to multiplying the neighborhood of each pixel by a different set of weights. This model reduces to a convolution operation if the degradation kernel is spatially-invariant, i.e., the same weights are used for the neighborhoods of all pixels. The reason to consider the general case of a spatially-variant and time-dependent kernel is due to the full processing chain of satellite images and their multitemporal nature. For satellite MISR, it is common practice to operate on orthorectified images, both due to data availability and because this simplifies the image registration process. However, the orthorectification process distorts the original sensor grid and interpolates pixel values and thus can be seen a further degradation which is spatially-variant due to the grid resampling. Furthermore, since the satellite attitude changes over time, the orthorectification process, and hence the degradation kernel, is different for each time instant. Besides the degradation due to the orthorectification process, the degradation kernel also includes the effect of the optics point spread function (PSF) which is space- and time-invariant.

Since the degradation kernel depends on both the optical system and the image processing chain, its realization can be highly dependent on the satellite under consideration and change over time. In order for the self-supervised training procedure to be effective and generalize beyond the training set, it is imperative to estimate $\mathbf{K}_t$ accurately and precisely. In fact, an accurate estimation for the given satellite and time instant allows us to invert its effects as best as possibile. However, we also need to characterize the expected variability in the $\mathbf{K}_t$ values. Training by simulating all the possible variations in $\mathbf{K}_t$ realizations allows the MISR model to become robust to new unseen realizations (if sufficiently similar to the training ones).

In theory, knowing the instrument PSF and the parameters of the orthorectification process should allow to analytically estimate $\mathbf{K}_t$ to some extent. However, in practice, the scarce availability of such information calls for blind estimation methods from the images themselves. Some kernel estimation methods [13] based on neural networks have recently

enjoyed some success but they are currently limited to single-image problems.

4. EXPERIMENTAL RESULTS

In this section we perform a series of experiments that aim at verifying the potential for the development of unsupervised MISR techniques in remote sensing. We test the performance of a few unsupervised model with respect to the state-of-the-art in remote sensing MISR and validate the model presented in the previous section. For this analysis, we use the Proba-V dataset for both training and testing. Unsupervised 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.

Our baseline study utilizes the state-of-the-art PIUnet neural network architecture [2], whose performance under supervised training is well known. We then study a few handcrafted degradation kernels in the aforementioned unsupervised training process in order to assess their impact on the SR performance. This can be considered a baseline study as we do not modify the original architecture nor we introduce other trainable components, which could improve the performance of unsupervised SR. Indeed, future work will focus on a joint model integrating estimation of the true degradation kernel (rather than a handcrafted definition) and of the SR image. For all experiments, we use 10 LR images, where 9 are degraded to CR using the kernel under investigation, downsampled by a factor of 3 and provided as network input, and the remaining LR image serves as target for the loss function. We use the original values for hyperparameters and the L1 loss for training for both the unsupervised and supervised PIUnet. The results are therefore compatible with the well-studied supervised setting.

Multiple degradation kernels were tested. First, bicubic interpolation was used as it is a widespread choice in absence of kernel information, albeit we will see that it leads to poor generalization. For all the other cases, we assumed a degradation model with anisotropic Gaussian filters [15]. We experimented with the use of a single spatially-invariant anisotropic Gaussian filter with a random covariance matrix for each image. The generated kernels have size 9×9 and a covariance matrix with random eigenvalues $\sigma_1, \sigma_2 \sim \mathcal{U}(0.5, 1.5)$ and rotation angle $\theta \sim \mathcal{U}(0, \frac{\pi}{2})$. Finally, we tested the more general case in which the kernel changes pixel-by-pixel and time-by-time. Two variants of this model: i) the filters used for every pixel are independent from each other, each following the same distribution equal to the one of the spatially-invariant case; ii) a handcrafted a spatial correlation pattern between the filters used in neighbouring pixels using the following equations:

$$\sigma_{1,2}(u,v) = \alpha_{1,2}\frac{\sigma_{1,2}(u-1,v) + \sigma_{1,2}(u,v-1)}{2}, \quad (2)$$

$$\alpha_{1,2} \sim \mathcal{U}(0.9, 1.1)$$

$$\theta(u,v) = \frac{\theta(u-1,v) + \theta(u,v-1)}{2} + \beta, \quad (3)$$

$$\beta \sim \mathcal{U}\left(-\frac{\pi}{8}, \frac{\pi}{8}\right)$$

with the value of $\sigma_{1,2}(0,0)$ being initialized using $\sigma_1, \sigma_2 \sim \mathcal{U}(0.5, 1.5)$ and $\theta \sim \mathcal{U}(0, \frac{\pi}{2})$. This latter case models the idea that a process like orthorectification is expected to produce a smoothly-varying spatial degradation.

The quality of the SR image compared to the HR was estimated using the corrected PSNR (cPSNR) metric [16]. Table 4 reports some quantitative results of the unsupervised and supervised approaches, as well as of a few classical model-based methods. It can be noticed that, despite the lack of HR training images, unsupervised deep learning approaches outperform traditional methods (bicubic+mean and IBP [14]). It can also be noticed that using the bicubic kernel as degradation model for training leads to poor generalization and the resulting cPSNR is only marginally better than using IBP. Substantial gains can be obtained by proper modeling of the degradation kernel. In particular, for the NIR dataset, the best performance is achieved by using a space- and time-variant declination of the degradation, suggesting that the most general formulation of the degradation model in Eq. (1) is indeed needed. These results suggest that unsupervised MISR for satellite images has potential to be close to supervised training in performance, while not requiring images beyond the LR observations. Indeed, since the presented results relied on handcrafted unoptimized degradation models, it is expected that end-to-end architectures for MISR that properly estimate both the degradation kernel and the SR image could exhibit strong performance, even compared to supervised approaches, while solving the data availability and domain gap issues. Future work will pursue such avenue.

5. CONCLUSIONS

This paper presented a preliminary study on the topic of unsupervised satellite super-resolution from multi-temporal images, motivated by the difficulty of collecting high-resolution imagery beyond the native sensor resolution. We showed that the problem requires careful modelling of the degradation process incurred by the observed LR images, realizing its time- and space-varying nature. However, our preliminary analysis with handcrafted degradation models showed that it is indeed possible to train deep learning models that significantly outperform classic unsupervised approaches and are close to supervised training. This result shows promise and the next step to bridge the gap to supervised training will focus on the development of a single architecture that can di-

Table 1. Quantitative performance - cPSNR (dB)

	Classic		Unsupervised deep SR				Supervised deep SR
	Bicubic int. +Mean	IBP [14]	PIUnet w/ Bicubic kernel	PIUnet w/ SI kernel	PIUnet w/ Pixel kernel (uncorr.)	PIUnet w/ Pixel kernel (corr.)	PIUnet
cPSNR (NIR)	45.44	45.96	46.08	46.78	46.69	46.98	48.41
cPSNR (RED)	47.34	48.21	48.24	49.02	48.99	48.97	50.53

rectly estimate the degradations and use it both for training and inference. We also would like to encourage increased public availability of non-orthorectified imagery in order to properly study its effects on satellite MISR.

6. REFERENCES

[1] 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.

[2] 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.

[3] Tai An, Xin Zhang, Chunlei Huo, Bin Xue, Lingfeng Wang, and Chunhong Pan, "Tr-misr: Multiimage super-resolution based on feature fusion with transformers," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1373–1388, 2022.

[4] 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.

[5] 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.

[6] Brian B. Moser, Federico Raue, Stanislav Frolov, Sebastian Palacio, Jörn Hees, and Andreas Dengel, "Hitchhiker's guide to super-resolution: Introduction and recent advances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–21, 2023.

[7] Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Kishor Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch, "Direct unsupervised super-resolution using generative adversarial network (dusgan) for real-world data," *IEEE Transactions on Image Processing*, vol. 30, pp. 8251–8264, 2021.

[8] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu, "Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 998–1008.

[9] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte, "Deep burst super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9209–9218.

[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] 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.

[12] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani, "Blind super-resolution kernel estimation using an internal-gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[13] 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.

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

[15] Gernot Riegler, Samuel Schulter, Matthias Rüther, and Horst Bischof, "Conditioned regression models for non-blind single image super-resolution," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 522–530.

[16] 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.