Multi-Image Super-Resolution for Remote Sensing using Deep Recurrent Networks

Md Rifat Arefin[1,2] Vincent Michalski[1,2] Pierre-Luc St-Charles[1] Alfredo Kalaitzis[3]
Sookyung Kim[4] Samira E. Kahou[1,5] Yoshua Bengio[1,2]

[1]Mila - Quebec AI Institute [2]Université de Montréal [3]Element AI
[4]Lawrence Livermore National Laboratory [5]McGill University

rifat.arefin515@gmail.com
{michals, pierreluc.stcharles, ebrahims, yoshua.bengio}@mila.quebec
freddie@element.ai
kim79@llnl.gov

Abstract

High-resolution satellite imagery is critical for various earth observation applications related to environment monitoring, geoscience, forecasting, and land use analysis. However, the acquisition cost of such high-quality imagery due to the scarcity of providers and needs for high-frequency revisits restricts its accessibility in many fields. In this work, we present a data-driven, multi-image super resolution approach to alleviate these problems. Our approach is based on an end-to-end deep neural network that consists of an encoder, a fusion module, and a decoder. The encoder extracts co-registered highly efficient feature representations from low-resolution images of a scene. A Gated Recurrent Unit (GRU)-based module acts as the fusion module, aggregating features into a combined representation. Finally, a decoder reconstructs the super-resolved image. The proposed model is evaluated on the PROBA-V dataset released in a recent competition held by the European Space Agency. Our results show that it performs among the top contenders and offers a new practical solution for real-world applications.

1. Introduction

Enhancing the quality of aerial and satellite imagery is one of the most prominent and challenging problems in remote sensing. The processes behind the acquisition pipelines that determine the quality of the imagery rely heavily on the quality of the sensors themselves, whether electro-optical, radar, or laser-based. High-precision sensors which can observe the earth with high resolution have historically been very expensive despite regular advances in technology over the past few decades. Recently, the availability of low-cost, low-resolution imagery provided by commercial space industries is exploding due to the deployment of new satellite constellations [39]. While these offer very interesting perspectives in terms of revisit frequency and coverage, there is a detachment between their capabilities and the current need for high-resolution imagery.

In this work, we present a super-resolution method based on a neural architecture to enhance the quality of satellite imagery. Our data-driven technique can potentially help bridge the gap between the needs for low-cost and high-resolution data using already existing acquisition capabilities. We approach the super-resolution problem from the image reconstruction perspective which aims at generating a high-resolution image based on one or more low-resolution images. Super-resolution in the special case of remote sensing has two important particularities that we can highlight:

1. **Variety of data sources.** Although high-resolution data sources are expensive, these can still be used in parsimony for the validation of enhancement methods. On the other hand, low-resolution data are plentiful and sometimes even publicly available at no cost. However, these may involve different acquisition sources, locations, or times, and may thus require special care in the way they are combined for super-

resolution purposes.

2. **Data misalignment and pixel-level inconsistencies.** Satellite imaging systems typically orbit the earth with pre-defined speeds and paths. Pixel-level inconsistencies can however still occur even with a well-calibrated system, and sub-pixel registration is often a necessary processing step for applications using several images at once. On the other hand, evolving land use and climate phenomena can have an impact on the similarity between images of the same region even if calibration issues are taken into account.

Modern super-resolution solutions have to take these considerations into account in order to achieve practical results.

Recent advances in deep learning have offered numerous new avenues for the single-image super-resolution (SISR) problem [8, 9, 37, 41, 26]. Most works related to SISR rely on neural networks to learn complex hierarchical representations for different versions of an observed region at varying resolutions. These representations are used to reconstruct high-quality images from low-quality ones based on the high-level understanding of the observed scene's underlying phenomena. The one-to-one mapping (in terms of inputs-to-ouputs) of the SISR approach may however hinder the quality of the reconstruction due to the low amount of information present in the original image. As a result, SISR solutions may lead to the "hallucination" of high-resolution details based on their training bias. In such cases, fake patterns with no discernible link to the input image may appear in the observed scene depending on the requested output resolution. This possibility may be cause for concern in applications that could lead to the misguidance of scientists or decision makers. A simple solution to this problem is to increase the amount of input information by instead using multiple low-resolution images at once. This technique is called multi-image super-resolution (MISR). The challenge for MISR approaches then becomes information fusion (or registration) due to the noisy nature of the imaging process. In general, MISR is capable of more accurate high-resolution reconstructions than SISR as it aggregates more information extracted from multiple views of the target region [22].

Image enhancement with the MISR approach is especially well suited for remote sensing for two major reasons. First, as mentioned earlier, it is fairly easy to obtain multiple low-resolution satellite images and a handful of high-resolution ones for the proper evaluation of the reconstruction of specific regions. Second, the use of multiple low-resolution images can alleviate some of the issues caused by misalignments and pixel-level inconsistencies through information fusion. In reality, atmospheric turbulence and sensor noise may even help regularize the information fusion process in a natural way.

Our solution tackles the MISR problem in an end-to-end fashion using a neural network-based model which we call *MISR-GRU*. Our processing pipeline is composed of three stages: we first encode the input low-resolution images into a set of low-resolution feature maps. Then, these are fused using a recurrent structure to obtain a combined scene representation. Finally, we decode this scene representation into a single high-resolution image. In the first stage, we choose a reference image to implicitly co-register the low-resolution information in the encoding space. In the second stage, we use a Convolutional Gated Recurrent Unit (ConvGRU) [1] model to fuse multiple low-resolution feature maps by capturing within-and-across-view relationships. In the third stage, we employ deconvolution layers [48] to reconstruct a high-resolution image from the fused feature maps. Our contributions mainly lie in the unified end-to-end architecture we propose. More specifically:

1. We formulate a feature encoding strategy which extracts effective features and acts as an implicit co-registration of low-resolution (LR) views in the embedding space.

2. We propose an efficient information fusion approach based on a Recurrent Neural Network (RNN) architecture: ConvGRU. This architecture can examine correlations between features within-and-across views obtained from a variable-length sequence of input low-resolution images.

3. We evaluate our solution on real-world imagery collected by the European Space Agency (ESA)'s PROBA-V satellite. Our results demonstrate the proposed architecture's competitive performance on the ESA's Advanced Concepts Competition portal, Kelvins[1], in comparison with other state-of-the-art solutions.

The rest of the paper is organized as follows: Section 2 reviews relevant works and image super-resolution techniques, Section 3 describes the proposed *MISR-GRU* model architecture, Section 4 presents our results as obtained from the Kelvins portal, and Section 5 concludes with an overview.

2. Related Work

Single Image Super Resolution. The recent advances in deep learning have provided a considerable number of new ideas to tackle the super-resolution problem. One of the early models for SISR was proposed by Dong et al. [8]: they proposed using an approximate mapping from low-resolution features to high-resolution ones based on three-layer Convolutional Neural Networks (CNNs). Motivated

[1] https://kelvins.esa.int/

by this work, several other authors proposed to adapt other deep learning architectures like RNNs [40, 10, 28], Residual CNNs [17], and GANs [15] to tackle SISR. Among them, Ledig et al. [26] introduced a ResNet-inspired architecture, "SRResNet", which preserves batch normalization inside the original residual blocks. This allows their model to require significantly less memory and allows the adaptation of several ideas introduced for image de-blurring [33]. Similarly, Lim et al. [29] proposed a new type of residual block by removing batch normalization and proposing their own type of residual (or "skip") connection. Skip connections in general are quite beneficial for CNNs as they constrain a layer to only learn the residual between its input and output [17]. The DenseNet architecture proposed by Huang et al. [18] is also based on this idea, and was adapted for SISR by Tong et al. [42]. More recently, Gao et al. [14] proposed Multi-scale Deep Neural Networks (Ms-DNNs) to tackle high-resolution image reconstruction when the up-scaling factors of image pairs are unknown and different from each other. In their solution, they reconstruct the details of an image by employing Multi-scale Residual Networks (MsRNs) in the downscaling spaces based on the residual blocks.

Multi Image Super Resolution. As detailed earlier, MISR approaches aim to reconstruct hidden high-resolution details using multiple low-resolution observations of the same scene. In contrast with SISR which has been extensively studied in the deep learning literature, MISR has seen slower progress due to the need to address the additional challenges of co-registration and information fusion. MISR faces the fundamental problem of de-aliasing, i.e. disentangling high-frequency artifacts caused by low-resolution sampling under varying phases [43]. The seminal work of Tsai et al. [43] reformulated this task through the co-registration of low-resolution images in the frequency domain. Over the years, other image registration and data fusion techniques [20, 12, 2] were introduced that also focused on the reconstruction of high-frequency details lost in the image acquisition process. In addition to these problems, MISR approaches must often address almost-random inconsistencies in images that increase the complexity of their registration. These inconsistencies include, for example, geometric distortions, blur, and pixel noise. Results can be improved by assuming that prior knowledge on these issues are being used at evaluation time [36]. Many modern optimization-based approaches to MISR build a generative model that, given a high-resolution image, simulates the acquisition of low-resolution images. An initial guess for the high-resolution image is then improved by minimization of the error between simulated and ground-truth low-resolution images. To reduce the parameter search space and to derive objective functions, these methods commonly model additive noise and prior knowledge about the acquisition process explicitly, using e.g. Tikhonov regularization [34], Huber potential [36] or Total Variation [11, 45]. When image degradation processes are too complex or unknown, non-parametric strategies become more interesting. High-resolution details can be for example reconstructed using a patch-based approach. These details can be inferred using sparse coding or sparse dictionary methods [23, 49, 47], or simply using k-nearest neighbor search [13, 3].

Recently, deep learning techniques have been applied to address the MISR problem in an end-to-end learning framework. Kawulok et al. [22] proposed a deep learning method for video image super-resolution for natural images which apply shift-and-add strategy for fusion without any image registration. For satellite image MISR of PROBA-V competition, DeepSUM [30] method is proposed which uses a convolutional neural network (CNN) architecture to register and fuse multiple unregistered LR images from the same scene. Moreover, another solution named HighResnet [7] from the same competition uses a CNN-based model to solve registration problem implicitly and merge the LR view in a recursive way. We briefly summarize some methods of interest that we compare our results to in Section 4.2.

3. Model

We formulate our proposed architecture and model around the task of MISR for satellite imagery for applications related to land cover and vegetation growth monitoring. Our primary goal is to offer a solution to the issue of the multiple-to-one mapping of MISR and support a variable number of LR images as input. We also wish to offer an intelligent end-to-end design that can tackle various types of pixel-level inconsistencies automatically. Accordingly, this section introduces our *MISR-GRU* architecture.

3.1. Problem Formulation

Let us denote by θ, α, and β the parameters of our Encoder, Fusion, and Decoder modules (respectively), and γ is the (up-)scaling factor. We define the i^{th} LR image of a scene as $l_i \in R^{c \times h \times w}$. Here, c, h, and w are the (channel-wise) depth, height, and width of the input LR image, respectively. For our evaluation, c is 1 and γ is 3. We denote the ground-truth high-resolution (HR) image for the same scene as $H \in \mathbb{R}^{c \times \gamma h \times \gamma w}$. The predicted output of our model is denoted by $\hat{H} = F^{\gamma}_{(\theta,\alpha,\beta)}(\{l_1, \ldots, l_K\})$, where K is the number of LR views of a scene.

An overview of our model is shown in Figure 1. It can be split into three modules (left to right): 1) *Encoder*, which encodes relevant features from the low-resolution images into latent representations; 2) *Fusion Module*, which merges the latent representations across the set of input images; and 3) *Decoder*, which reconstructs the target high-resolution

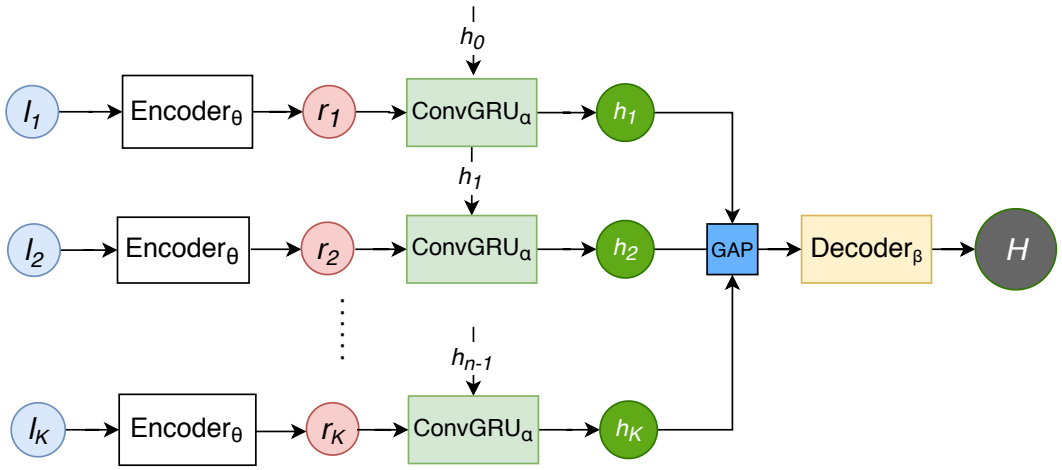

Figure 1. Architectural overview of the proposed MISR-GRU.

image. Each module is described in the following subsections.

3.2. Encoder

MISR approaches assume that the low-resolution image sets that are provided as input contain more information than any of such image alone. To exploit this information, the model has to be robust against uncontrolled factors such as blurriness or distortions. In our specific problem formulation, we do not assume that the images of a common scene are already co-registered. Deep learning approaches have been introduced already to tackle the generic image registration problem with respect to a reference image [27]. We follow a similar strategy and use a *reference* image, noted $q \in R^{c \times h \times w}$, as an auxiliary representation of our inputs. This reference image is obtained by computing the pixel-wise median value of several input images in a way similar to the technique of [7]. We concatenate the intermediate feature representations of *reference* image with the feature representations of each of the LR images so that the network figures out the co-registration implicitly with respect to the *reference* image representations.

We show in Figure 2 a symbolic view of our *Encoder* module.

Given the input tensors $(l_i)_{i=1}^K$, and q, the network is trained to produce feature representations, denoted by $(r_i)_{i=1}^K$. To obtain these representations, we first embed l_i and q using Unit1. This produces two feature maps, $l_i^{'} \in R^{64 \times h \times w}$ and $q' \in R^{64 \times h \times w}$ respectively. Then $l_i^{'}$ is concatenated with q' and passed through Unit2. The result is a co-registered feature map $r_i \in R^{64 \times h \times w}$. The Unit1 and Unit2 consist of two convolutional layers and two residual blocks [17] each. This type of block layout is inspired by [21] and combines two convolutional layers and activation functions without any batch-normalization layer. For

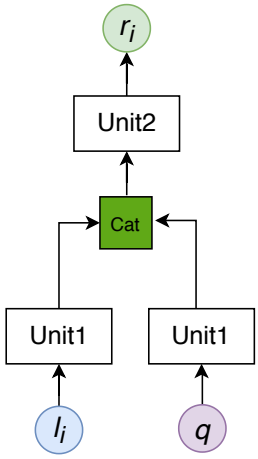

Figure 2. Overview of an encoder module.

all the convolutions, we use 3x3 kernels with feature map of 64, and for the activations, we use Parametric ReLU [16]. Formally, the feature representations $(r_i)_{i=1}^K$ are computed using

$$q' = \text{Unit1}(q), \tag{1}$$
$$l'_i = \text{Unit1}(l_i), \tag{2}$$
$$r_i = \text{Unit2}([l'_i, q']), \tag{3}$$

where we represent the concatenation operation between two tensors A and B as $[A, B]$.

3.3. Fusion Module

As mentioned earlier, many image fusion techniques for MISR rely on weighted average [4, 25] or directly apply CNNs on the stacked images [19]. In contrast, RNNs are

generally designed for modeling sequential data. Although our problem is not truly sequential in nature, we can still consider the set of input low-resolution images as a sequence. A direct advantage of this strategy is the possibility of processing sets of any size, while still being able to model within-and-across-view relationships between images of the input set. The ConvGRU model [1] is an extension of the GRU block introduced in [5] in which fully-connected layers are replaced with convolutional layers both in the input-to-state and state-to-state connections. The hidden state $\mathbf{h}_t$ of the ConvGRU is recurrently connected to its sequential neighbors. It is updated by convolution from the input feature map $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ as follows:

$$
\begin{aligned}
u_t &= \sigma(W_i * [x_t, h_{t-1}] + b_i) \\
r_t &= \sigma(W_j * [x_t, h_{t-1}] + b_j) \\
c_t &= tanh(W_h * [x_t, r_t \odot h_{t-1}] + b_h \\
h_t &= (1 - u_t) \odot h_{t-1} + u_t \odot c_t
\end{aligned}
$$

Here, $W_.$ and $b_.$ are convolution kernels and biases respectively. $*$ and $\odot$ are the convolution and element-wise multiplication operators, respectively.

Each convolutional transition is defined using 2-D kernels that represent the receptive fields of h_t's hidden units in the input x_t and in the previous hidden state h_{t-1}.

In formal terms, ConvGRU acts as the fusion network $f_\alpha : \mathbb{R}^{K\times 64\times h\times w} \rightarrow \mathbb{R}^{K\times C_{h^{\text{GRU}}}\times h\times w}$ that converts K input representations $(r_i)_{i=1}^{K}$ to K hidden representations $(\mathbf{h_i^{\text{GRU}}})_{\mathbf{i=1}}^{\mathbf{K}}$. Then, global average pooling (GAP) is used on the sequence dimension to return a fixed-size representation $\mathbf{h}_{\text{avg}} \in \mathbb{R}^{C_{h^{GRU}}\times W\times H}$. The fusion network f_α can be stacked as necessary to obtain more complex features that cover a wider receptive field in the input images. In that case, we denote the embeddings for L levels of convolution layers as $\mathbf{h}^{\text{GRU}^{(l)}}$ with $l = 1, \cdots, L$. In any case, the K different representations $(r_i)_{i=1}^{K}$ obtained from the encoder are used to initialize the first level of the fusion network. More specifically,

$$
(\mathbf{h_i})_{\mathbf{i=1}}^{\mathbf{K}} = (\mathbf{h_i^{\text{GRU}^{(L)}}})_{\mathbf{i=1}}^{\mathbf{K}} = f_\alpha(\mathbf{r}_1, \cdots, \mathbf{r}_K) \tag{4}
$$

$$
\mathbf{h}_{\text{avg}} = \text{GAP}(h_1, \ldots, h_K) \tag{5}
$$

For our own experiments, we set L to 2 and $C_{h^{GRU}}$ to 64.

3.4. Decoder

The role of the decoder is simply to upsample the combined representations $\mathbf{h}_{\text{avg}}$ obtained by averaging the ConvGRU hidden states into a tensor $\hat{H}$ the size of the target high-resolution image H. Formally, $\hat{H}$ is computed via

$$
\hat{H} = \text{decoder}_\beta^\gamma(\mathbf{h}_{\text{avg}}) \in \mathbb{R}^{c\times\gamma h\times\gamma w}. \tag{6}
$$

The decoding network itself consists of a deconvolution layer [48] and a 1×1 convolution layer. The deconvolution layer is useful here as it increases the accuracy and the training speed of this final stage of our proposed architecture, as reported in [9, 38, 44]. We used 3×3 kernels, 64 features maps, a stride of 3, and Parametric ReLU as the activation function.

Finally, an ultimate 1×1 convolution layer is used to project the decoder's output feature map of dimensions $\mathbb{R}^{64\times\gamma h\times\gamma w}$ into the target image space of dimension $\mathbb{R}^{c\times\gamma h\times\gamma w}$.

3.5. Target Image Registration

The super-resolution image $\hat{H}$ generated by the proposed *MISR-GRU* model may be shifted with respect to the ground-truth high-resolution image H due to misalignments in the image calibration process. Training without trying to account for this issue results in blurry reconstructions. To compensate for this issue, we rely on the *ShiftNet* strategy of [7] which is inspired by the idea of Deep *HomographyNet*[6]. In short, *ShiftNet* warps the predicted image based on deep learning based estimation of translation parameters so that it fits with the target high-resolution image during training. At evaluation time, the direct output of the MISR-GRU model is reconstructed as-is without warping.

3.6. Loss Function

The typical way to formulate the super-resolution training objective is to minimize the reconstruction error between the target high-resolution image and the model's prediction. In this spirit, the Mean Squared Error (MSE) is commonly used in practice due to its interpretability and effectiveness [45]. Measuring the Peak Signal-to-Noise Ratio (PSNR) can also give a good idea of a model's performance. However, both the PSNR and MSE are sensitive to biases in brightness. In concordance with the evaluation guidelines of the challenge dataset (detailed in the next section), we opt to use a corrected metric for our loss function. We settle on a variant of the MSE called the corrected MSE (cMSE) which equalizes the brightness in both predicted and target images. The cMSE is defined by the following equations

$$
b = \frac{1}{|S|}\sum\Big((H - \hat{H})\cdot S\Big), \tag{7}
$$

$$
\text{cMSE}(H, \hat{H}) = \frac{1}{|S|}\sum(H - \hat{H} + b)^2, \tag{8}
$$

where b is the brightness bias (a scalar), S is the binary map which indicates clear pixels in H, and where the images H and $\hat{H}$ are both normalized to the $[0, 1]$ range. We can then calculate the corrected PSNR (cPSNR) from the cMSE using

$$\text{cPSNR}(H, \hat{H}) = -10 \times \log_{10}(\text{cMSE}(H, \hat{H})). \quad (9)$$

Finally, to define a loss function to minimize, we use the negative cPSNR:

$$\text{loss}(H, \hat{H}) = -\text{cPSNR}(H, \hat{H}) \quad (10)$$

4. Results

Our training and evaluation dataset is composed of images collected by the PROBA-V earth observation satellite. The dataset was released as part of a super-resolution competition held by ESA's Advanced Concepts Team in 2019 [31]. The imagery consists of red and Near InfraRed (NIR) spectral bands with 14-bit depth. The low-resolution images were prepared with a shape of 128×128 pixels while the high-resolution (target) images contained 384×384 pixels, meaning the (up-)scaling factor γ is exactly 3. In terms of ground resolution, each pixel corresponds to 300m and 100m in the low-resolution and high-resolution images, respectively. The image patches are provided with a binary mask that indicates the quality/status of each pixel. This mask identifies areas with clouds, water, shadows, or other uncontrolled pixel-wise inconsistencies. The dataset contains a total of 1450 target regions; of these, 1160 were preselected for training and 290 were reserved for testing. Each region is covered with an average of 19 low-resolution images, with a minimum of 9. An example of low-resolution patches, reconstructed SR image by *MISR-GRU*, and a high-resolution crop is shown in Figure 3.

4.1. Performance Score Calculation

The full methodology for the calculation of a method's performance score is detailed on the Kelvins portal[2]. This methodology slightly compensates for image misalignments by looking for the best super-resolution results under a set of limited 2D pixel shifts $\langle u, v \rangle \in \mathcal{C}$. The methodology also returns a normalized score that relates how well a method performs in comparison to a simple baseline (based on bi-cubic interpolation). Formally, given $\text{cPSNR}^*(H)$ the cPSNR obtained by the baseline for a high-resolution image H, the score of a method is given as

$$z(\hat{H}) = \max_{\langle u,v \rangle \in \mathcal{C}} \frac{\text{cPSNR}^*(H)}{\text{cPSNR}(H_{\langle u,v \rangle}, \hat{H})}, \quad (11)$$

where $H_{\langle u,v \rangle}$ is the target high-resolution image H shifted in 2D space by $\langle u, v \rangle$ pixels. Lastly, the overall score of a submission is obtained by averaging all its $z(\hat{H})$ scores over all test regions.

[2] https://kelvins.esa.int/proba-v-super-resolution/scoring/

4.2. Experimental Setup

Our model was implemented in PyTorch [35] and made publicly available[3]. The model was optimized end-to-end using Adam [24] starting with an initial learning rate of 0.0007 and gradual learning rate decay with a factor of 0.97 whenever the validation score plateaued for more than 2 epochs. We trained the model for 400 epochs with a batch size of 16. The training process took roughly 13 hours on a NVIDIA Titan RTX with memory of 24GB. During inference our model can super-resolve around 14 scenes per second in the same GPU if each scene is processed individually without batching. Because of the memory constraint and to improve generalization by data augmentation, we trained our model with randomly cropped 64×64 LR and corresponding 192×192 HR patches. As our model is fully convolutional, at test time we feed full LR images of spatial size 128×128 as input. *MISR-GRU* has around 0.9M parameters. Table 2 gives an overview of the architecture.

We compare our results with those of numerous methods taken from the literature or proposed by the ESA for the competition. These are briefly summarized below:

- **ESA Baseline:** computes the average of most clear low-resolution images and upsamples the result using bicubic interpolation.
- **FSRCNN-8 [9]**: independently computes a high-resolution version of 8 low-resolution inputs using FSRCNN and averages them into a final prediction.
- **SRResNet-1 [26]**: finds the most clear low-resolution input image based on binary status map which indicates the clear pixels and upsamples it using the *SRResNet* based SISR approach.
- **SRResNet-1 + ShiftNet**: same as above, but registers the most clear low-resolution image with the ground-truth during training using *ShiftNet*.
- **SRResNet-6 + ShiftNet**: same as above, but applied to six independent low-resolution images. The results are also averaged into a final prediction.
- **DeepSUM [32]**: this is actually a variant of the *SRResNet-6 + ShiftNet* approach where several low-resolution images are independently upsampled and gradually merged back into the final prediction.
- **ACT [30]**: concatenates 5 low-resolution images channel-wise as input and uses a CNN to upsample them directly.
- **HighResNet [7]**: co-registers up-to 32 low-resolution images with a reference image and learns feature representations that are recursively fused together. The

[3] https://github.com/rarefin/MISR-GRU

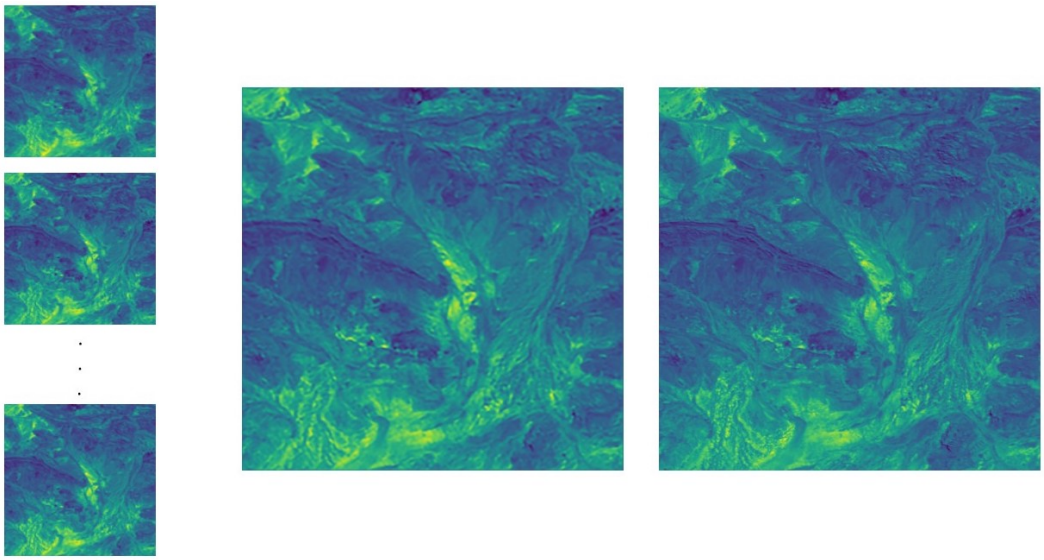

Figure 3. Example of a candidate region for super-resolution: a) overlapping low-resolution input images, b) reconstructed super resolution image by *MISR-GRU*, c) target high-resolution image.

Table 1. PROBA-V super-resolution evaluation scores

Method	Public score
SRResNet-1 [26]	1.0095
SRResNet-1 + ShiftNet	1.0002
ESA Baseline	1.0000
FSRCNN-8 [9]	0.9927
ACT [30]	0.9874
SRResNet-6 + ShiftNet [46]	0.9808
ConvGRU-24 (ours)	0.9808
ConvGRU-9 + ShiftNet (ours)	0.9776
ConvGRU-13 + ShiftNet (ours)	0.9757
ConvGRU-24 + ShiftNet (ours)	0.9581
ConvGRU-24* + ShiftNet (ours)	0.9502
HighResNet [7]	0.9496
DeepSUM [32]	0.9488
MISR-GRU (ours)	**0.9484**

high-resolution image is generated and aligned using *ShiftNet*.

The performance of the model depends on the number of LR views, co-registration of them with respect to the reference view, and registration of reconstructed super-resolved image. The model should also ideally be permutation invariant, i.e. the fusion result should not depend on the order in which LR images are processed. To prevent overfitting and to encourage permutation invariance, we randomly select a subset of LR images. For configurations *ConvGRU-24 + ShiftNet*, *ConvGRU-24* + ShiftNet* and *MISR-GRU* we use the random selection introduced in [7], which prefers those views with low amounts of occlusion. For other configurations, all LR views are equally likely to be drawn. We summarize the different network architecture configurations used in our experiments below:

- **ConvGRU-24**: uses a sequence of up-to 24 low-resolution images as input and reconstructs a high-resolution image from the last hidden state representation of the ConvGRU.
- **ConvGRU-9 + ShiftNet**: same as *ConvGRU-24* using 9 input images, but also registers the reconstructed output with the target image during training using *ShiftNet*.
- **ConvGRU-13 + ShiftNet**: same as *ConvGRU-9 + ShiftNet* using 13 input images.
- **ConvGRU-24 + ShiftNet**: same as *ConvGRU-24* except that it also uses *ShiftNet* for registration of the output with the target and the random selection of LR views as in [7].
- **ConvGRU-24* + ShiftNet**: same as *ConvGRU-24 + ShiftNet*, but in addition using the co-registration with a reference image described in Section 3.2.
- **MISR-GRU**: described in Section 3 uses up-to 24 LR images.

Table 2. MISR-GRU Configuration & Parameters

Modules	Sub-modules	Blocks	Layers	Input Channels	Output Configuration (kernel, filters, stride, padding)	Number of Parameters
Encoder	Unit1		Conv2d PReLU	1	3×3, 64, s1, p1	640 1
		ResidualBlocks	[Conv2d PReLU Conv2d PReLU] × 2	[64 64] × 2	[3×3, 64, s1, p1 3×3, 64, s1, p1] × 2	[36928 1 36928 1] × 2
			Conv2d	64	3×3, 64, s1, p1	36928
	Unit2		Conv2d PReLU	128	3×3,64,s1,p1	73792 1
		ResidualBlocks	[Conv2d PReLU Conv2d PReLU] × 2	[64 64] × 2	[3×3, 64, s1, p1 3×3, 64, s1, p1] × 2	[36928 1 36928 1] × 2
			Conv2d	64	3×3, 64, s1, p1	36928
Fusion Module	ConvGRU Levels		[Conv2d Conv2d Conv2d] × 2	[128 128 128] × 2	[3×3, 64, s1, p1 3×3, 64, s1, p1 3×3, 64, s1, p1] × 2	[73792 73792 73792] × 2
Decoder			ConvTranspose2d PReLU Conv2d	64 64	3×3, 64, s3, p0 1×1, 1, s1, p0	36928 1 65
						923468 (total)

4.3. Discussion

Table 1 lists the PROBA-V super-resolution challenge performance scores for our different configurations as well as the competing and baseline solutions.

We can first observe that the use of a registration technique in the super-resolution process is quite beneficial across the views. The co-registration of LR images implicitly with respect to a reference image and registration of reconstructed super resolved images with respect to the ground-truths before calculating loss significantly improves overall performance. We also found that use of more LR images is beneficial (but we observed that the images need to generally be as clear as possible). For example, using 13 images as input seemed to beat the configuration that used 9 in combination with *ShiftNet*, however, more LR images don't help, if the super resolved image is not registered (*ConvGRU-24*). *MISR-GRU* make use of up-to 24 low-resolution images and implicitly co-register them, *ConvGRU* based fusion strategy and registration of super-resolution image using *ShiftNet* and performs as one of the top scorers.

We also observed in that if images were affected by unmasked inconsistencies such as clouds, smoke, or shadows, the reconstruction process was often misdirected. Incorporating metadata in the form of acquisition parameters (view angle, latitude, longitude, etc.) could help learn more invariant feature representations. In the future, we will work on addressing these issues.

5. Conclusion

We introduced a new deep learning-based MISR technique that relies on the implicit co-registration of feature maps of low-resolution images to produce high-quality output images in an end-to-end fashion. Our approach is based on a convolutional RNN fusion architecture that can aggregate an arbitrary number of low-resolution images into a combined representation for decoding. We also employed *ShiftNet* during the training process to bootstrap our model's registration capabilities. We believe this approach offers a good practical solution to address the trade-off problem between image quality and acquisition cost/complexity in remote sensing applications.

Acknowledgements

We thank Ishaan Kumar, Zhichao Lin, Kris Sankaran, Julien Cornebise, Anthony Ortiz and Jason Jo for useful discussions. We are also thankful to the Advanced Concepts Team of the ESA for organizing the Kelvin competition.

References

[1] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015. 2, 5

[2] David Capel and Andrew Zisserman. Super-resolution from multiple views using learnt image models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001. 3

[3] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 3

[4] Peter Cheeseman, Bob Kanefsky, Richard Kraft, John Stutz, and Robin Hanson. Super-resolved surface reconstruction from multiple images. In *Maximum Entropy and Bayesian Methods*, pages 293–308. 1996. 4

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*, 2014. 5

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv:1606.03798*, 2016. 5

[7] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020. 3, 4, 5, 6, 7

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Patt. Anal. Machine Intellig.*, 38(2):295–307, 2015. 2

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016. 2, 5, 6, 7

[10] David Eigen, Jason Rolfe, Rob Fergus, and Yann LeCun. Understanding deep architectures using a recursive convolutional network. *arXiv:1312.1847*, 2013. 3

[11] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE Trans. Image Proc.*, 13(10):1327–1344, 2004. 3

[12] J Michael Fitzpatrick, Derek LG Hill, Calvin R Maurer, et al. Image registration. *Handbook of medical imaging*, 2:447–513, 2000. 3

[13] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, (2):56–65, 2002. 3

[14] Shangqi Gao and Xiahai Zhuang. Multi-scale deep neural networks for real image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1026–1034, 2015. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 4

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 3

[19] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European Conference on Computer Vision*, pages 353–369, 2016. 4

[20] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991. 3

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016. 4

[22] Michal Kawulok, Pawel Benecki, Krzysztof Hrynczenko, Daniel Kostrzewa, Szymon Piechaczek, Jakub Nalepa, and Bogdan Smolka. Deep learning for fast super-resolution reconstruction from multiple images. In *Real-Time Image Processing and Deep Learning 2019*, 2019. 2, 3

[23] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[25] M Prema Kumar and P Rajesh Kumar. Pixel level weighted averaging technique for enhanced image fusion in mammography. *Intl. J. Inf. Electron. Eng*, 3(5), 2015. 4

[26] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2, 3, 6, 7

[27] Hongming Li and Yong Fan. Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv:1709.00799*, 2017. 4

[28] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3367–3375, 2015. 3

[29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 3

[30] Marcus Märtens, Dario Izzo, Andrej Krzic, and Daniël Cox. Super-resolution of proba-v images using convolutional neural networks. *Astrodynamics*, 3(4):387–402, 2019. 3, 6, 7

[31] Marcus Märtens, Dario Izzo, Andrej Krzic, and Daniël Cox. Super-resolution of proba-v images using convolutional neural networks. *Astrodynamics*, 3(4):387–402, 2019. 6

[32] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *arXiv:1907.06490*, 2019. 6, 7

[33] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017. 3

[34] Nhat Nguyen, Peyman Milanfar, and Gene Golub. A computationally efficient superresolution image reconstruction algorithm. *IEEE Trans. Image Proc.*, 10(4):573–583, 2001. 3

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[36] Lyndsey C Pickup, Stephen J Roberts, and Andrew Zisserman. Optimizing and learning for super-resolution. In *British Machine Vision Conference*, volume 9, pages 4–7, 2006. 3

[37] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image upscaling with super-resolution forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3799, 2015. 2

[38] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 5

[39] Lake A Singh, William R Whittecar, Marc D DiPrinzio, Jonathan D Herman, Matthew P Ferringer, and Patrick M Reed. Low cost satellite constellations for nearly continuous global coverage. *Nature Communications*, 11(1):1–7, 2020. 1

[40] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Y Ng. Convolutional-recursive deep learning for 3d object classification. In *Advances in Neural Information Processing Systems*, pages 656–664, 2012. 3

[41] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3147–3155, 2017. 2

[42] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4799–4807, 2017. 3

[43] Roger Y. Tsai and Thomas S. Huang. Multi-frame image restoration and registration. 1984. 3

[44] Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. End-to-end image super-resolution via deep and shallow convolutional networks. *IEEE Access*, 7:31959–31970, 2019. 5

[45] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *arXiv:1902.06068*, 2019. 3, 5

[46] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9127–9135, 2018. 7

[47] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Proc.*, 19(11):2861–2873, 2010. 3

[48] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2528–2535, 2010. 2, 5

[49] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730, 2010. 3